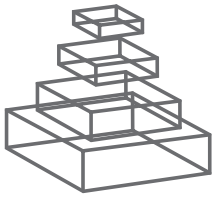
frontiers
RESEARCH TOPICS

INTERFACES BETWEEN LANGUAGE AND COGNITION

Topic Editors
Yury Y. Shtyrov, Andriy Myachykov and Christoph Scheepers

frontiers in PSYCHOLOGY



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-147-5
DOI 10.3389/978-2-88919-147-5

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

INTERFACES BETWEEN LANGUAGE AND COGNITION

Topic Editors:
Yury Y. Shtyrov, Medical Research Council (MRC), United Kingdom
Andriy Myachykov, Northumbria University, United Kingdom
Christoph Scheepers, University of Glasgow, United Kingdom

Cognitive mechanisms underlying linguistic communication do not only rely upon retrieval and processing of linguistic information; they also involve constant updating and organizing of this linguistic information in relation with other, more general, cognitive mechanisms. Some existing theoretical models assume such a tight interactive link between domain-general and domain-specific sources of information in the cognitive organization of the linguistic faculty and during language use. Domain-specific constraints may include, for example, grammatical as well as lexical and pragmatic knowledge. Domain-general constraints comprise processing limitations imposed by the cognitive mechanisms of memory, attention, learning, and social interaction. However, much of the existing research tends to focus on one or the other of the aforementioned areas, while integrative accounts are still rather sparse at present. Therefore, the aim of this Research Topic of Frontiers in Cognition is to bring together researchers who, with in their respective research fields and by using different methodologies, represent integrative approaches to the study of language. We invite submissions from a wide range of interrelated areas of research: cognitive architectures of language, aspects of language processing, linguistic development, bilingualism, language embodiment, neuropsychology of linguistic function, among others. We would like to solicit original research contributions discussing behavioral, neurophysiological, and computational evidence as well as papers on methodological and/or theoretical aspects of the interplay between linguistic and non-linguistic cognitive processes.

Table of Contents

frontiers in
PSYCHOLOGY

EDITORIAL
published: 06 May 2013
doi: 10.3389/fpsyg.2013.00258

Interfaces between language and cognition

***Andriy Myachykov*[1]*, *Christoph Scheepers*[2] *and Yury Y. Shtyrov*[3]**

[1] *Department of Psychology, Northumbria University, Newcastle upon Tyne, UK*
[2] *Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK*
[3] *MRC Cognition and Brain Sciences Unit, Cambridge University, Cambridge, UK*
**Correspondence: andriy.myachykov@northumbria.ac.uk*

Edited by:
Bernhard Hommel, Leiden University, Netherlands

Reviewed by:
Bernhard Hommel, Leiden University, Netherlands

One of the most intriguing and challenging questions in the interdisciplinary study of mental processes and underlying brain mechanisms is how language is related to thought. The question is by no means new. Scholars have attempted to unravel the relationship between language and thought since the early days of Western philosophy. Recent theories range from strictly modular accounts of linguistic processing to fully integrated theories, according to which linguistic processes strongly interact with more general cognitive mechanisms such as attention, memory, and action control. Unfortunately, theoretical exchange between proponents of these different views is often lacking. In part, this is due to the interdisciplinary nature of the question itself. As a result, researchers representing various disciplines often fail to engage in an exchange of theoretical views, research ideas, and methodological expertise. The present Frontiers' Special Topic provides a platform for such dialogue. It features contributions discussing the latest advances and challenges in the frontline research on language and cognition and attempts to provide a joint discussion forum for a wide range of researchers from the domains of cognitive psychology, neuroscience, and psycholinguistics, among others. These researchers follow different theoretical approaches and use different experimental methodologies. What unites them is their goal to understand the mechanisms underlying the interplay between linguistic and general cognitive processes.

General cognitive mechanisms in linguistic communication do not only include retrieval and processing of linguistic information; they also rely upon constant updating and organizing of this linguistic information in relation with other, more general representations. Some existing theoretical models assume a tight interactive coupling between domain-general and domain-specific sources of information in the cognitive organization of the linguistic faculty. Domain-specific constraints may include, for example, grammatical as well as lexical and pragmatic knowledge. Domain-general constraints comprise processing limitations imposed by the cognitive mechanisms of memory, attention, learning, and social interaction. However, much of the existing research tends to focus on one or the other of the aforementioned areas, while integrative accounts are still rather sparse at present. The aim of this Special Topic of Frontiers in Cognition is therefore to bring together researchers who, within their respective research fields and by using different methodologies, represent integrative approaches to the study of language. Our Research Topic presents a collection of seventeen excellent articles that include original research, commentaries, opinions, and reviews.

A number of papers in this Topic discuss neurophysiological and behavioral evidence about the interface between language, perception, and attention. Research discussed by Roelofs and Piai (2011) suggests that word planning does not always require full executive attention while specific attention deficits may contribute to impaired language performance. The results discussed by Meyer and colleagues (2012) demonstrate how gaze shifts can be linked to the process of phonological encoding with specific focus on word production automaticity. The article by Myachykov et al. (2012) presents evidence about the special role attention plays in determining the assignment of grammatical roles and the associated syntactic choice in visually situated sentence production. Papers by Huettig et al. (2012), Knoeferle et al. (2011), and Kaiser (2012) provide complementary evidence about the involvement of the language-cognition interface during sentence comprehension in visually situated contexts. The contribution by Shtyrov (2011) reports novel findings about rapid pre-attentive mapping of novel word forms, as evidenced by changes in the dynamics of brain responses within very short exposures. Finally, Hussey and Novick (2012) report intriguing evidence about the benefits of executive control training for grammatical processing in ambiguous contexts.

The question of coordination between interlocutors during dialogue is raised in two articles. Gambi and Pickering (2011) used a novel interactive methodology in order to demonstrate that interlocutors constantly coordinate their sentences to represent their partner's knowledge. They then use these representations to build unfolding predictions, which they take into account when planning self-generated utterances. Similarly, Dale and colleagues (2011) use eye-movement synchronization between interlocutors as evidence for rapid approximation of actions in dialogue and the emergence of a single coordinated interactive system.

Three papers in our Topic discuss embodied and grounded aspects of language processing. Lupyan (2012) addresses the question of the language-cognition interplay from the point of view of how language affects cognition and perception. In particular, Lupyan (2012) reviews evidence showing that performance on tasks that have been presumed to be non-verbal is rapidly modulated by language. Klemfuss et al. (2012) discuss effects of language on perception by critically reviewing evidence suggesting top-down influences of linguistic representations on visual

feature detection. Their own research suggests that visual search is disrupted by the automatic activation of irrelevant linguistic representations. Another important aspect of the grounded view of language is the role played by perception and action systems in the organization of abstract knowledge. Scorolli et al. (2011) discuss the crucial role played by embodied theories of cognition in linguistic experience for abstract words.

A number of papers in this Special Topic discuss architectural properties of the language-cognition interface. For example, Menenti et al. (2012) investigated how brain areas adapt to repetition of various sentence properties, thereby unraveling the neuronal infrastructure for the specific components of semantic encoding. Mashal and colleagues (2012) present a novel cortical network model for observation and imitation of speech. Their results show that the network models for observation and imitation comprise the same essential structure but differ in important features that reflect distinct connectivity patterns. Andric and Small (2012) contribute to the debate by discussing how the brain processes language and co-occurring gestures. Finally, Naylor et al. (2012) focus on cognitive and electrophysiological correlates of the bilingual Stroop effect by analysing corresponding ERP components in bilingual speakers. Their research shows, among other things that color words from both languages created response conflict and that the between-within language Stroop effect reflects complex brain activity with contributions from language both and color at different task points.

REFERENCES

Andric, M., and Small, S. L. (2012). Gesture's neural language. *Front. Psychol.* 3:99. doi: 10.3389/fpsyg.2012.00099

Dale, R., Kirkham, N. Z., and Richardson, D. C. (2011). The dynamics of reference and shared visual attention. *Front. Psychol.* 2:355. doi: 10.3389/fpsyg.2011.00355

Gambi, C., and Pickering, M. J. (2011). A cognitive architecture for the coordination of utterances. *Front. Psychol.* 2:275. doi: 10.3389/fpsyg.2011.00275

Huettig, F., Mishra, R. K., and Olivers, C. N. L. (2012). Mechanisms and representations of language-mediated visual attention. *Front. Psychol.* 2:394. doi: 10.3389/fpsyg.2011.00394

Hussey, E. K., and Novick, J. M. (2012). The benefits of executive control training and the implications for language processing. *Front. Psychol.* 3:158. doi: 10.3389/fpsyg.2012.00158

Kaiser, E. (2012). Taking action: a cross-modal investigation of discourse-level representations. *Front. Psychol.* 3:156. doi: 10.3389/fpsyg.2012.00156

Klemfuss, N., Prinzmetal, W., and Ivry, R. B. (2012). How does language change perception: a cautionary note. *Front. Psychol.* 3:78. doi: 10.3389/fpsyg.2012.00078

Knoeferle, P., Carminati, M. N., Abashidze, D., and Essig, K. (2011). Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension. *Front. Psychol.* 2:376. doi: 10.3389/fpsyg.2011.00376

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Front. Psychol.* 3:54. doi: 10.3389/fpsyg.2012.00054

Mashal, N., Solodkin, A., Dick, A. S., Chen, E. E., and Small, S. L. (2012). A network model of observation and imitation of speech. *Front. Psychol.* 3:84. doi: 10.3389/fpsyg.2012.00084

Menenti, L., Petersson, K. M., and Hagoort, P. (2012). From reference to sense: how the brain encodes meaning for speaking. *Front. Psychol.* 2:384. doi: 10.3389/fpsyg.2011.00384

Meyer, A. S., Wheeldon, L., van der Meulen, F., and Konopka, A. (2012). Effects of speech rate and practice on the allocation of visual attention in multiple object naming. *Front. Psychol.* 3:39. doi: 10.3389/fpsyg.2012.00039

Myachykov, A., Thompson, D., Garrod, S., and Scheepers, C. (2012). Referential and visual cues to structural choice in visually situated sentence production. *Front. Psychol.* 2:396. doi: 10.3389/fpsyg.2011.00396

Naylor, L. J., Stanley, E. M., and Wicha, N. Y. Y. (2012). Cognitive and electrophysiological correlates of the bilingual stroop effect. *Front. Psychol.* 3:81. doi: 10.3389/fpsyg.2012.00081

Roelofs, A., and Piai, V. (2011). Attention demands of spoken word planning: a review. *Front. Psychol.* 2:307. doi: 10.3389/fpsyg.2011.00307

Scorolli, C., Binkofski, F., Buccino, G., Nicoletti, R., Riggio, L., and Borghi, A. M. (2011). Abstract and concrete sentences, embodiment, and languages. *Front. Psychol.* 2:227. doi: 10.3389/fpsyg.2011.00227

Shtyrov, Y. (2011). Fast mapping of novel word forms traced neurophysiologically. *Front. Psychol.* 2:340. doi: 10.3389/fpsyg.2011.00340

Received: 18 April 2013; accepted: 18 April 2013; published online: 06 May 2013.

*Citation: Myachykov A, Scheepers C and Shtyrov YY (2013) Interfaces between language and cognition. Front. Psychol. **4**:258. doi: 10.3389/fpsyg.2013.00258*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 07 November 2011
doi: 10.3389/fpsyg.2011.00307

Attention demands of spoken word planning: a review

Ardi Roelofs* and Vitória Piai

Centre for Cognition, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Daniel Kleinman, University of California San Diego, USA
Pia Knoeferle, Bielefeld University, Germany

****Correspondence:***
Ardi Roelofs, Centre for Cognition, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Spinoza Building B.01.08, Montessorilaan 3, 6525 HR Nijmegen, Netherlands.
e-mail: a.roelofs@donders.ru.nl

Attention and language are among the most intensively researched abilities in the cognitive neurosciences, but the relation between these abilities has largely been neglected. There is increasing evidence, however, that linguistic processes, such as those underlying the planning of words, cannot proceed without paying some form of attention. Here, we review evidence that word planning requires some but not full attention. The evidence comes from chronometric studies of word planning in picture naming and word reading under divided attention conditions. It is generally assumed that the central attention demands of a process are indexed by the extent that the process delays the performance of a concurrent unrelated task. The studies measured the speed and accuracy of linguistic and non-linguistic responding as well as eye gaze durations reflecting the allocation of attention. First, empirical evidence indicates that in several task situations, processes up to and including phonological encoding in word planning delay, or are delayed by, the performance of concurrent unrelated non-linguistic tasks. These findings suggest that word planning requires central attention. Second, empirical evidence indicates that conflicts in word planning may be resolved while concurrently performing an unrelated non-linguistic task, making a task decision, or making a go/no-go decision. These findings suggest that word planning does not require full central attention. We outline a computationally implemented theory of attention and word planning, and describe at various points the outcomes of computer simulations that demonstrate the utility of the theory in accounting for the key findings. Finally, we indicate how attention deficits may contribute to impaired language performance, such as in individuals with specific language impairment.

Keywords: attention, dual-task performance, naming, reading, response times, specific language impairment

INTRODUCTION

In his classic monograph *Die Sprache*, Wundt (1900) – the founder of modern scientific psychology and psycholinguistics – criticized the now classic model of normal and impaired word production and comprehension of Wernicke (1874) by arguing that processing words is an attention demanding rather than an automatic process. According to Wundt (1900), a central attention system located in the frontal lobes of the human brain actively controls a lexical network centered around perisylvian brain areas, described by the Wernicke model. More than a century later, attention and language are among the most intensively researched abilities in the cognitive neurosciences, but the relation between these abilities has largely been neglected. Modern computational models of normal and impaired picture naming and word reading build in many respects on Wernicke's model (e.g., Dell et al., 1997; Coltheart et al., 2001), but they do not address Wundt's concern of how word processing is controlled by attention. Word processing in these models makes no demands on non-linguistic processing mechanisms or resources and does not depend on top-down attentional control.

There is increasing evidence, however, that most language processes underlying picture naming and word reading cannot proceed without paying some form of attention. It is generally assumed that the central attention demands of a process are indexed by the extent to which the process delays the performance of a concurrent unrelated task (e.g., Johnston et al., 1995). Circumstantial evidence that language performance requires central attention is provided by the effort associated with talking or reading in a foreign language or talking while driving a car in heavy traffic. Experiments on dual-task performance provide evidence that the alleged prototype of an automatic language process, the generation of a phonological code (e.g., Ferreira and Pashler, 2002), in fact requires central attention in both word reading (Reynolds and Besner, 2006) and picture naming (Roelofs, 2008a). The evidence asks for a reexamination of the century-old dogma that most processes in naming and reading are automatic (i.e., require no attentional capacity), which is the aim of the present article.

As Wundt (1900) argued, understanding the relation between attention and language is of great theoretical and practical importance. To the extent that central attention determines language performance, psycholinguistic models that only address language processes are incomplete. Moreover, evidence suggests that inefficient allocation and deficits of attention contribute to language impairments in aphasia and dyslexia (e.g., Murray, 1999; Shaywitz and Shaywitz, 2008). Also, there is evidence that attention deficits play a role in the impaired language performance of individuals with specific language impairment (SLI; e.g., Im-Bolter et al., 2006; Spaulding et al., 2008; Finneran et al., 2009). A better understanding of the relation between attention and language may help improve therapeutic interventions.

Attention comprises several different abilities. A prominent theory proposed by Posner and colleagues (e.g., Posner and Raichle, 1994; Posner and Rothbart, 2007) distinguishes three fundamental aspects, referred to as alerting, orienting, and executive control. Alerting concerns the achievement and maintenance of an alert state. This maintenance is often referred to as sustained attention or vigilance. Orienting concerns the direction of processing toward a location in space by overtly shifting gaze or covertly shifting the locus of processing while keeping the eyes fixed. Executive control concerns the regulative processes that ensure that thoughts and actions are in accordance with goals. This ability is engaged in the selection among competitors, controlled memory retrieval, the coordination of processes, and the allocation of central attentional capacity (e.g., Baddeley, 1996). Executive control also regulates overt and covert orienting. The performance of the central executive depends on the state of vigilance (e.g., Kahneman, 1973). In the present article, we concentrate on the executive control aspect of attention, and briefly address the orienting of attention (i.e., gaze shifting) and aspects of sustained attention.

The remainder of the article is organized as follows. We start by outlining a computationally implemented theory of attention and word planning, which serves as the theoretical framework for the present article. The theory acknowledges many aspects of the work of Wernicke, but also addresses Wundt's critique by including assumptions on how word planning is controlled. Next, we review empirical results indicating that in several task situations, processes up to and including phonological encoding in word planning delay, or are delayed by, the performance of concurrent unrelated non-linguistic tasks. These findings suggest that word planning requires central attention. Then, we review empirical results indicating that conflicts in word planning may be resolved while concurrently performing an unrelated non-linguistic task, making a task decision, or making a go/no-go decision. These findings suggest that word planning does not require full central attention, contrary to claims in the literature that processes in word planning cannot occur in parallel with processes in non-linguistic tasks if both require central attention (e.g., Ferreira and Pashler, 2002; Dell'Acqua et al., 2007; Ayora et al., 2011). At various points, we describe the outcomes of computer simulations that demonstrate the utility of our theory in accounting for the key empirical findings on word production under divided attention conditions. We end by indicating how attention deficits may contribute to impaired language performance, such as in individuals with SLI.

OUTLINE OF A THEORY OF ATTENTION IN WORD PLANNING

FUNCTIONAL ASPECTS

In the present article, attention to word planning is addressed using the theoretical framework of the WEAVER++ model (Roelofs, 1992, 1997, 2003, 2004, 2006, 2007, 2008a,b,c; Levelt et al., 1999; Piai et al., 2011). This model makes a distinction between declarative (i.e., associative memory) and procedural (i.e., rule system) aspects of word planning (cf. Ullman, 2004). Information about words is stored in a large associative network. WEAVER++'s lexical network is accessed by spreading activation while condition–action rules determine what is done with the activated lexical information depending on the goal (e.g., to name a picture or read aloud a word). When a goal is placed in working memory, processing in the system is focused on those rules that include the goal among their conditions. The rules mediate attentional influences by selectively enhancing the activation of target nodes in the network in order to achieve speeded and accurate picture naming and word reading.

A fragment of the lexical network of WEAVER++ is illustrated in **Figure 1**. According to the model, the naming of pictures involves the activation of nodes for lexical concepts, lemmas, morphemes, phonemes, and syllable motor programs in associative memory. The nodes are selected by condition–action rules. For example, naming a pictured dog involves the activation and selection of the representation of the concept DOG(X), the lemma of *dog* specifying that the word is a noun (for languages such as Dutch, lemmas also specify grammatical gender), the morpheme ⟨dog⟩, the phonemes/d/,/ɔ/, and/g/, and the motor program [dɔg]. Not shown is that lemmas also allow for the specification of morphosyntactic parameters, such as number (singular, plural) for nouns and number, person (first, second, third), and tense (past, present) for verbs, so that condition–action rules can retrieve appropriate inflectional morphemes (e.g., plural or past tense endings). Activation spreads from level to level, whereby each node sends a proportion of its activation to connected nodes. Consequently, network activation induced by perceived pictures decreases with network distance. The activation flow from concepts to phonological forms is limited unless attentional enhancements are involved to boost the activation of target concept nodes.

The model assumes that perceived pictures have direct access to concepts [e.g., DOG(X)] and only indirect access to word forms (e.g., ⟨dog⟩ and/d/,/ɔ/,/g/), whereas perceived words have direct access to word forms and only indirect access to concepts. Consequently, naming pictures requires concept selection, whereas words can be read aloud without concept selection. The latter is achieved by mapping input word forms (e.g., the visual word DOG) directly onto output word forms (e.g., ⟨dog⟩ and/d/,/ɔ/,/g/), without engaging concepts and lemmas. With such direct form-to-form mapping, activation has to travel a much shorter network distance from input to output than with a mapping via concepts and lemmas. In word reading through the form-to-form route, the activation of target morphemes is enhanced by the attention system. Given the shorter network distance for word reading than picture naming, the attentional enhancements may be less for reading than naming, and successful reading relies much less on the enhancement than does naming.

As already explained, the activation enhancements in WEAVER++ are regulated by a system of condition–action rules. When a goal is placed in working memory, word planning is controlled by those rules that include the goal among their conditions. The activation enhancements are required until appropriate motor programs have been activated sufficiently, that is, above an availability threshold. The central executive determines how strongly and for how long the enhancement occurs. The required duration of the enhancement is assessed by monitoring the progress on word planning (i.e., the updating in working memory of subgoals to retrieve lemmas, morphemes, and so forth).

In planning words while simultaneously performing another task, the central executive coordinates the processes involved in

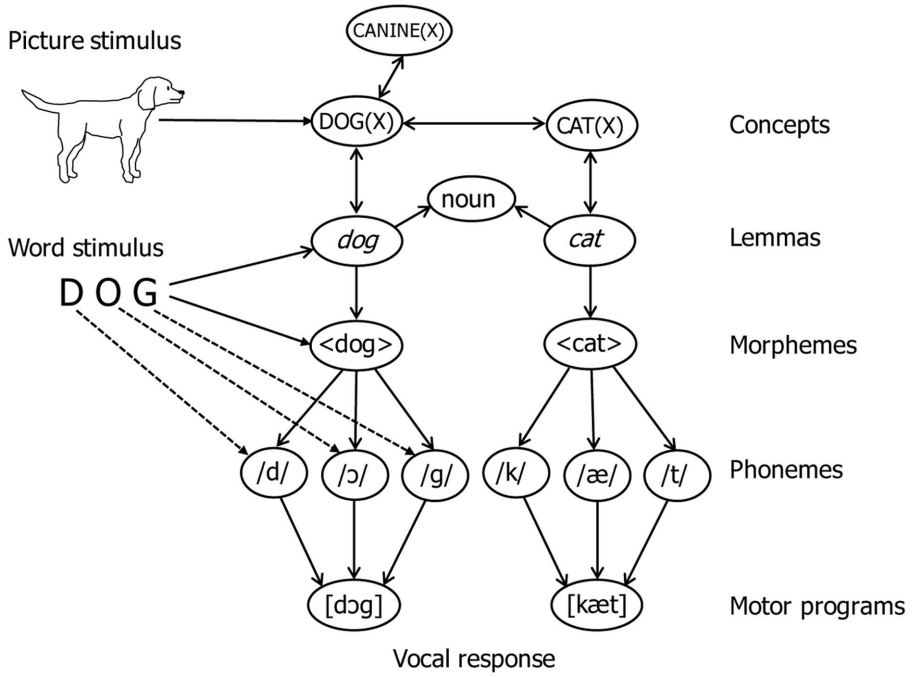


FIGURE 1 | Illustration of the lexical network of WEAVER++. Perceived pictures (e.g., of a dog) directly activate concept nodes and perceived words (e.g., DOG) directly activate lemma, morpheme, and phoneme nodes, after which other nodes become activated through spreading activation. The dashed lines indicate grapheme-to-phoneme correspondences.

such a way as to maintain acceptable levels of speed and accuracy, to minimize resource consumption and crosstalk between tasks, and to satisfy instructions about task priorities (cf. Meyer and Kieras, 1997a). Resources include the buffering of input, throughput, or output representations (e.g., motor programs) and central attentional capacity. The model assumes that attentional capacity is limited (i.e., there is a limit to the top-down activation enhancements), but the limit depends on the effort exerted at any time. The degree of effort depends on the demand of the concurrent processes, which is evaluated during task performance (cf. Kahneman, 1973).

NEURAL ASPECTS

To assess the neural basis of the word planning process, Indefrey and Levelt (2004) conducted a meta-analysis of 82 neuroimaging studies on word production. The meta-analysis included picture naming (e.g., say "dog" to a picture of a dog), word generation (producing a use for a noun, e.g., say "walk" to the word DOG), word reading (e.g., say "dog" to the word DOG), and pseudoword reading (e.g., say "doz" to DOZ). Pseudowords are letter strings that include only combinations of letters that are permissible in the spelling of a language and that are pronounceable for speakers of the language. According to the meta-analysis, percepts and concepts in picture naming are activated in occipital and inferiotemporal regions of the brain. The middle part of the left middle temporal gyrus seems to be involved in lemma retrieval. Next, activation spreads to Wernicke's area, where morphemes of the word seem to be retrieved. Activation is then transmitted to Broca's area for morphological assembly as well as phoneme processing and syllabification (i.e., phonological encoding), see also Sahin et al. (2009) and Ullman (2004), among others. Next, motor programs are accessed. The sensorimotor areas control articulation. The form-to-form mapping in word reading may be accomplished by activating occipital and inferiotemporal regions (i.e., the occipito-temporal sulcus) for orthographic processing, inferioparietal cortex and the areas of Wernicke and Broca for aspects of form encoding, and motor areas for articulation (cf. Shaywitz and Shaywitz, 2008; Dehaene, 2009).

Neuroimaging studies have shown that especially the anterior cingulate cortex (ACC) and lateral prefrontal cortex (LPFC) are implicated in the executive control aspect of attention to word planning. For example, the ACC and LPFC are more active in word generation (say "walk" to the word DOG) when the attention demand is high than in word reading (say "dog" to DOG) when the demand is much lower (Petersen et al., 1988; Thompson-Schill et al., 1997). The increased activity in the frontal areas disappears when word selection becomes easy after repeated generation of the same word (Petersen et al., 1998). Moreover, activity in the frontal areas is higher in picture naming when there are several good names for a picture, making selection difficult, than when there is only a single appropriate name (Kan and Thompson-Schill, 2004). Also, the frontal areas are more active when retrieval fails and words are on the tip of the tongue than when words are readily available (Maril et al., 2001). Frontal areas are also more active in naming pictures with semantically related words superimposed (e.g., naming a pictured dog combined with the word CAT) than without word distractors (e.g., a pictured dog combined with XXX), as observed by de Zubicaray et al. (2001). Thus, the

neuroimaging evidence suggests that medial and lateral prefrontal areas exert control over word planning. Along with the increased frontal activity, there is an elevation of activity in temporal areas for word planning (e.g., de Zubicaray et al., 2001).

Although both the ACC and LPFC are involved in executive control aspects of attention to word planning, these areas seem to play different roles. WEAVER++'s assumption that abstract condition–action rules mediate goal-oriented retrieval and selection processes in prefrontal cortex is supported by evidence from single cell recordings and hemodynamic neuroimaging studies (e.g., Sakai, 2008, for a review). Much evidence suggests that the dorsolateral prefrontal cortex is involved in maintaining goals in working memory (for a review, see Kane and Engle, 2002). Moreover, evidence suggests that the ventrolateral prefrontal cortex plays a role in selection among competing response alternatives (Thompson-Schill et al., 1997), the control of memory retrieval, or both (Badre et al., 2005). Researchers have found no agreement on whether the ACC performs conflict monitoring (e.g., Botvinick et al., 2001) or exerts regulatory influences over word planning processes, as has been assumed for WEAVER++ (Roelofs and Hagoort, 2002; Roelofs, 2003; Roelofs et al., 2006).

EVIDENCE THAT WORD PLANNING REQUIRES CENTRAL ATTENTION

CENTRAL ATTENTION DEMANDS OF PICTURE NAMING

The assumption that word planning requires attentional activation enhancements is not only supported by neuroimaging evidence, but also by evidence from chronometric studies. In a study by Roelofs et al. (2007), participants were shown pictures of objects (e.g., a dog) while hearing a tone or a spoken word presented 600 ms after picture onset. When a spoken word was presented (e.g., *desk* or *bell*), participants indicated whether it contained a pre-specified phoneme (e.g., /d/) by pressing a button. When the tone was presented, they indicated whether the picture name contained the phoneme (Experiment 1) or they named the picture (Experiments 2 and 3). Phoneme monitoring latencies for the spoken words were shorter when the picture name contained the pre-specified phoneme (e.g., dog – *desk*) compared to when it did not (e.g., dog – *bell*). However, no priming of phoneme monitoring was obtained when the pictures required no response but were only passively viewed (Experiment 4). Thus, passive picture viewing does not lead to significant phonological activation. These results suggest that attentional enhancements are a precondition for obtaining phonological activation from perceived pictures of objects.

In the passive-viewing condition of Roelofs et al. (2007), speakers may have paid some attention to the picture, but apparently not long enough to induce phonological activation. To assess how long attention needs to be sustained to a picture, eye movements and response times to the picture may be measured. Past research showed that while individuals can shift the focus of attention without an eye movement (covert orienting), they cannot move their eyes to one spatial location while paying full attention to another location (i.e., shifts of eye position require shifts of attention). Thus, a gaze shift (overt orienting) indexes a shift of attention (Wright and Ward, 2008). In a review of the literature on gazes and language performance, Griffin (2004) stated that "the production processes that appear to be resource demanding, based on dual-task performance, pupil dilation, and other measures of mental effort, are the same ones that are reflected in the duration of name-related gazes" (p. 222).

Research on spoken word planning has shown that speakers tend to gaze at words and pictures until the completion of phonological encoding (e.g., Meyer et al., 1998; Griffin, 2001; Korvorst et al., 2006). For example, when speakers are asked to name two spatially separated pictures (e.g., one on the left side of a computer screen and the other on the right side), they look longer at first-to-be-named pictures with disyllabic names (e.g., *baby*) than with monosyllabic names (e.g., *dog*) even when the picture recognition times are the same (Meyer et al., 2003). The effect of the phonological length suggests that the shift of gaze from one picture to the other is initiated only after the phonological form of the name for the picture has been encoded and the corresponding articulatory program is available. The executive control system appears to instruct the orienting system to shift gaze depending on the completion of phonological encoding. By making gaze shifts dependent on phonological encoding, resource consumption may be diminished. Articulating a word such as "dog" can easily take half a second or more. If gaze shifts are initiated as soon as the first picture is identified, the planning of the name for the second picture may be completed well before articulation of the name for the first picture has been finished. Consequently, the motor program of the second vocal response needs to be buffered for a relatively long time. By starting perception of the second picture only after the planning of the first picture name is completed sufficiently, the use of buffering resources can be limited. Another reason why gaze shifts are made dependent on the completion of phonological encoding is to reduce or prevent interference from the other picture name, which promotes the speed and accuracy of naming performance.

Malpass and Meyer (2010) provided evidence that the name of the second picture may interfere with planning the name of the first picture. The ease of naming the second picture was manipulated. Easy and difficult second pictures were matched for difficulty of picture recognition, but they differed in average naming latencies and error rates. Participants gazed longer at the first picture when the name of the second picture was easy than when it was more difficult to retrieve. This suggests that planning the name of the first picture suffers more interference from the easy than the difficult second pictures. However, when the processing of the first picture was made more difficult by presenting it upside down, no effect of second picture difficulty on the gaze duration for the first picture was found. These results suggest that participants can retrieve the names of foveated and parafoveal pictures in parallel, but only when the processing of the foveated picture does not demand too much attention.

The avoidance of response buffering and the prevention of interference from the second response are not the only reasons for a phonology-dependent gaze shift. Gaze shifts still depend on phonological encoding when the second naming response is replaced by a manual response to a left- or right-pointing arrow, so that there can be no interference from a second naming response (Roelofs, 2008a). Using the so-called psychological refractory period (PRP) procedure (cf. Pashler, 1998), speakers

were presented with pictures displayed on the left side of a computer screen and left- or right-pointing arrows displayed on the right side of the screen, as illustrated in **Figure 2**. The arrows ⟨ and ⟩ were flanked by two Xs on each side to prevent that they could be identified through parafoveal vision, which was the case for the second pictures in the study of Malpass and Meyer (2010). The picture and the arrow were presented simultaneously on the screen (SOA = 0 ms) or the arrow was presented 300 or 1000 ms after picture onset. The participants' tasks were to name the picture (Task 1) and to indicate the direction in which the arrow was pointing by pressing a left or right button (Task 2). Eye movements were recorded to determine the onset of the shift of gaze between the picture and the arrow. Phonological encoding was manipulated by having the speakers name the pictures in blocks of trials where the picture names shared the onset phoneme (e.g., *dog*, *doll*, *desk*), the homogeneous condition, or in blocks of trials where the picture names did not share the onset phoneme (e.g., *dog*, *bell*, *pin*), the heterogeneous condition. Earlier research has shown that picture naming RTs are smaller in the homogeneous than heterogeneous condition.

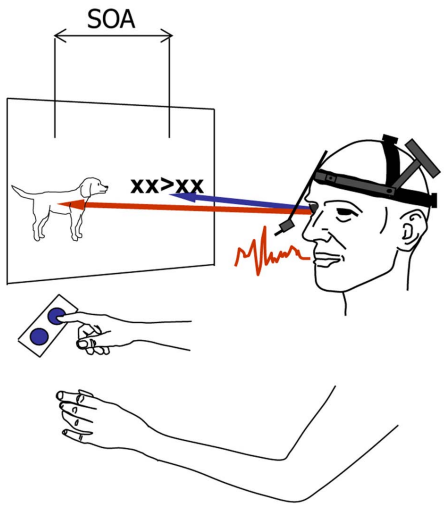


FIGURE 2 | Schematic illustration of the experimental set-up used in the eye tracking study of Roelofs (2008a). On each trial, participants named a picture and shifted their gaze to a left- or right-pointing arrow to manually indicate its direction by pressing a left or right button. Picture and arrow were presented at stimulus onset asynchronies (SOAs) of 0, 300, or 1000 ms. The latencies of vocal responding, gaze shifting, and manual responding were recorded.

Figure 3A shows the patterns of results. Phonological overlap in a block of trials reduced picture naming and gaze shifting latencies at all SOAs. Gaze shifts were dependent on phonological encoding even when they were postponed at the non-zero SOAs. Manual responses to the arrows were delayed and reflected the phonological effect at the short SOAs (i.e., 0 and 300 ms) but not at the long one (i.e., SOA = 1000 ms). These results suggest that gaze shifts still depend on phonological encoding when speakers name a picture and manually respond to an arrow. This finding suggests that the avoidance of response buffering and the prevention of interference from the second response are not the only reasons for a phonology-dependent gaze shift. Instead, some aspect of spoken word planning itself appears to be the critical factor. If attentional enhancements are required until the word has been planned far enough, this would explain why attention, indexed by eye gazes, is sustained to word planning until the phonological form is planned. This should hold regardless of the need for response buffering and the prevention of interference, as the eye tracking results indicate. **Figure 3B** shows the results of computer simulations of the experiment using WEAVER++, which we explain below.

To account for these results and related ones, the model assumes that participants decide which processes may run in parallel in Task 1 and Task 2 (i.e., how attention is divided). To this end, they set a point at which Task 2 processing is strategically suspended, called

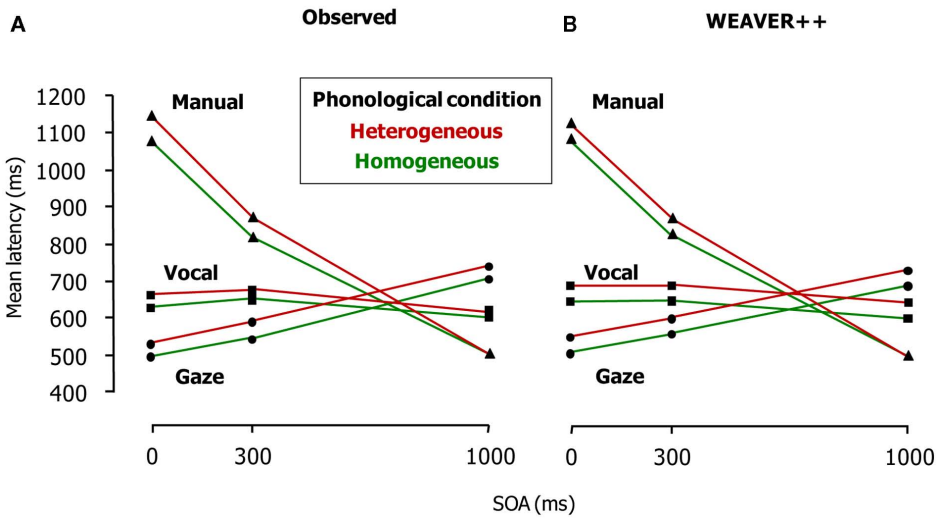


FIGURE 3 | Latencies of vocal responding (Task 1), gaze shifting, and manual responding (Task 2) in the dual-task study of Roelofs (2008a; Experiment 1). (A) Shows the real latency data of the eye tracking experiment and **(B)** shows the results of the computer simulations. SOA = stimulus onset asynchrony.

the "Task 2 lockout point" by Meyer and Kieras (1997b). Moreover, they set a criterion for when the shift of attention between Task 1 and Task 2 should occur. Reaching the shift criterion is called the occurrence of the "Task 1 unlocking event," which unlocks Task 2. The lockout point and shift criterion serve to maintain acceptable levels of speed and accuracy, to minimize resource consumption (including attentional capacity) and crosstalk between tasks, and to satisfy instructions about task priorities (i.e., the common instruction is that the Task 1 response should precede the Task 2 response). Presumably, the positions of the lockout point and shift criterion are determined on the basis of the initial trials of an experiment, when participants become familiar with the experimental situation, and the lockout point and criterion stay more or less constant throughout the experiment. At the beginning of each trial, the attention system enables both tasks, engages on Task 1 and temporarily suspends Task 2, instructs the ocular motor system to direct gaze toward the Task 1 stimulus, and maintains engagement on Task 1 and monitors performance until the task process reaches the task-shift criterion. Moreover, in homogeneous sets, the phonological encoder is instructed to prepare the phoneme that is shared by the responses in a set. Also during the planning of the target word, a saccade to the arrow is prepared. When the shift criterion is reached during the course of Task 1, attention disengages from Task 1 and shifts to Task 2, which is then resumed, directly followed by a signal to the saccadic control system to execute the prepared saccade to the Task 2 stimulus.

Figure 4 illustrates the timing of vocal responding, attention and gaze shifting, and manual responding in the model when the SOAs are short (i.e., 0 and 300 ms, Panel A) and long (i.e., 1000 ms, Panel B). The unlocking event corresponds to the completion of phonological encoding. At both short and long SOAs, the word planning reaches the unlocking event earlier in the homogeneous than the heterogeneous condition. This phonological facilitation effect is reflected in the naming and gaze shift latencies. Moreover, at short SOAs, the facilitation is reflected in the manual response latencies if there is a pause after the Task 2 lockout point. At the short SOAs, the pause is simply the waiting period until the eyes fixate the arrow so that it can be processed. Because gaze shifted earlier in the homogeneous than the heterogeneous condition, processing of the arrow (Task 2 stimulus) also started earlier in the homogeneous than the heterogeneous condition. Consequently, the phonological facilitation effect is reflected in the manual response RTs. However, at the long SOA, the phonological effect is reflected in the naming and gaze shift latencies, but not in the manual response latencies. This is because the phonological effect is absorbed when waiting for the arrow presentation. That is, at the long SOA, gaze has already shifted to the position on the screen where the arrow will later appear. If the arrow appears after the gaze has shifted in the heterogeneous condition, the processing of the arrow will start at the same moment in time for the homogeneous and heterogeneous conditions. Consequently, the phonological facilitation of vocal response planning will no longer be reflected in the manual response RTs. **Figure 3B** shows the WEAVER++ simulation results. A comparison with **Figure 3A** shows that the fit between model and data is good. The computer simulations demonstrate the utility of our theoretical account.

Further evidence that attention is sustained to word planning until the completion of phonological encoding comes from experiments by Cook and Meyer (2008) using the PRP procedure. Participants had to perform picture naming (Task 1) and manual tone discrimination (Task 2) tasks. In the critical conditions, the pictures were combined with phonologically related or unrelated distractors. Experiment 1 used distractor pictures, whereas the other experiments used distractor words, which were either clearly visible (Experiment 2) or masked (Experiment 3). Relative to the unrelated distractors, the phonologically related distractor

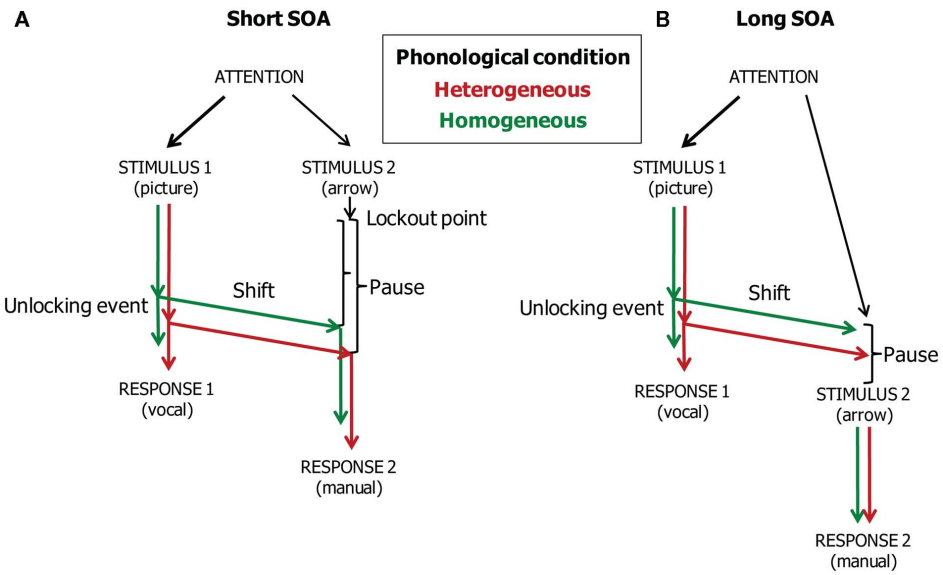


FIGURE 4 | Theoretical account of the time course of vocal responding (Task 1), attention and gaze shifting, and manual responding (Task 2) in the dual-task study of Roelofs (2008a). The stimulus onset asynchrony (SOA) is short **(A)** or long **(B)**.

pictures reduced the naming (Task 1) and manual (Task 2) RTs. Similarly, Roelofs (2008b) observed that the phonological effect of picture distractors on picture naming is present in the gaze durations. Cook and Meyer (2008) also obtained the phonological effect for the distractor words in picture naming, but only when the words were masked, not when they were clearly visible. The clearly visible distractor words yielded phonological facilitation in the naming RTs, but not in the manual RTs. For the manual RTs, the phonological effect tended to be one of interference (i.e., longer RTs on the related than unrelated trials) rather than facilitation. The presence of the phonological effect in the manual (Task 2) RTs for the picture and masked word distractors suggests that participants maintained attention to word planning in picture naming until the completion of phonological encoding. To explain the absence of a phonological facilitation effect in the manual RTs (or the presence of phonological interference) for the clearly visible word distractors, Cook and Meyer (2008) proposed that the phonological facilitation effect in picture naming was offset by longer self-monitoring durations in the phonologically related than unrelated condition.

Evidence from the study of Roelofs (2008a) supports the assumption of WEAVER++ that the allocation of attention in dual-task performance is not fixed but strategically determined (cf. Meyer and Kieras, 1997a). When speakers name pictures in homogeneous and heterogeneous trial blocks (Task 1) and manually respond to arrows or tones (Task 2), phonological encoding for word production delays the manual responses to the arrows (Roelofs, 2008a; Experiments 1–3) but not to the tones (Experiment 4). This suggests that speakers in the experiments of Roelofs (2008a) shifted attention earlier to the tones (i.e., before phonological encoding) than to the arrows (i.e., after phonological encoding).

Whereas (Roelofs, 2008a; Experiment 4) obtained no phonological effect in the tone task, Cook and Meyer (2008) observed a phonological effect on the response to the tones when Task 1 had picture distractors (Experiment 1) or masked word distractors (Experiment 3), whereas no phonological effect was obtained with visible word distractors (Experiment 2). These differences in results suggest that participants may set the shift criterion (i.e., when to shift attention to Task 2) differently depending on the exact circumstances. The shift criterion and lockout point are free parameters of the WEAVER++ model, but the parameter values are constrained. Evidence suggests that when Task 1 requires word planning, the shift criterion may differ in whether or not phonological encoding is completed before attention is shifted. If attention is shifted before phonological encoding, still some attentional capacity will have to be allocated to phonological encoding to make it possible. When Task 2 requires word planning, the lockout point may differ in whether or not lemma retrieval is completed before the planning process is suspended.

Evidence that attention may shift before phonological encoding was not only obtained by Roelofs (2008a; Experiment 4) and Cook and Meyer (2008; Experiment 2), but also by Ferreira and Pashler (2002). They had participants name the picture of picture–word combinations (Task 1) and indicate the pitch of a tone through button presses (Task 2). The SOAs between picture–word stimulus and tone were 50, 150, and 900 ms. The written distractor words were semantically related (e.g., pictured dog, distractor CAT), phonologically related (e.g., distractor DOLL), or unrelated to the picture names (e.g., distractor PIN). Compared to the unrelated distractor words, the semantically related words increased picture naming RTs and the phonologically related words reduced the RTs. Earlier research has suggested that the semantic interference arises in lemma retrieval, whereas the phonological facilitation arises in phonological encoding (cf. Levelt et al., 1999). Ferreira and Pashler (2002) observed that the semantic interference, but not the phonological facilitation, was propagated into the manual RTs. That is, the manual RTs were longer in the semantically related than unrelated condition, but equal in the phonologically related and unrelated conditions. These results suggest that attention was shifted from picture naming to tone discrimination before the onset of phonological encoding, in line with the results of the tone task obtained by Roelofs (2008a; Experiment 4).

Ferreira and Pashler (2002) observed that the semantic interference effect of word distractors in picture naming was carried forward to the manual RTs, suggesting that resolving the conflict underlying the interference requires attention. In line with this, Roelofs (2007) observed that participants gaze longer at picture–word stimuli in the semantically related than unrelated condition. Similarly, gaze durations depend on the amount of conflict in the color–word Stroop task (Roelofs, 2011). In a commonly used version of the Stroop task, participants name the color attribute of colored congruent or incongruent color–words (e.g., the words GREEN or RED in green ink, respectively; say "green") or neutral series of Xs. Naming RT is longer in the incongruent than in the neutral condition and often shorter in the congruent than in the neutral condition (for reviews, see MacLeod, 1991; Roelofs, 2003). In line with the RTs, participants gaze longer at incongruent than neutral stimuli and longer at neutral than congruent stimuli (Roelofs, 2011), which suggests that there are differences in attention demand among the Stroop conditions. Greater attentional effort is often reflected in a higher skin-conductance response, which is observed for the incongruent compared with the congruent Stroop condition (Naccache et al., 2005).

CENTRAL ATTENTION DEMANDS OF READING

It is often assumed that Stroop effects provide evidence for the automaticity of reading (e.g., MacLeod, 1991). The presence of interference and facilitation in this task is taken as evidence that participants automatically read the word, despite the instruction to ignore the word. However, given that the color and word are spatially integrated and part of one perceptual object (i.e., a colored word), it is also possible that Stroop effects reflect the difficulty of not allocating attention to the word in this task (cf. Kahneman, 1973; Pashler, 1998). On this view, word reading occurs in the Stroop task not because it happens automatically, but rather because the word inadvertently receives some of the attention that was meant for the color.

Accumulating evidence supports the attentional view of word reading in the Stroop task (e.g., Besner et al., 1997; Besner and Stolz, 1999). For example, when the color attribute of the color–word Stroop stimuli is removed (i.e., changed into neutral white color on a dark computer screen) 120 or 160 ms after stimulus presentation onset (e.g., RED in green ink is changed into RED in

neutral white ink), the magnitude of Stroop interference is reduced compared with the standard continuous presentation of the color until trial offset (La Heij et al., 2001). As argued by La Heij et al. (2001), the duration effect on Stroop interference is paradoxical: Whereas the only stimulus attribute present on the screen for most of the trial is an incongruent word, Stroop interference is less. The finding can be explained, however, if one assumes that removing the color attribute hampers the grouping of the color and word attributes into one perceptual object (i.e., a colored word) to which attention is allocated (cf. La Heij et al., 2001; Lamers and Roelofs, 2007). Because the written color–word receives less attention in the removed than in the continuous condition, the magnitude of the Stroop interference will also be less, as empirically observed. The utility of this account was demonstrated by computer simulations of the exposure duration effect using WEAVER++ (Roelofs and Lamers, 2007). Color removal not only reduces Stroop interference, but also Stroop facilitation. Moreover, color removal reduces gaze durations, suggesting reduced attention demand (Roelofs, 2011).

Whereas the findings on Stroop task performance suggest that word reading is affected by visual (input) attention, Reynolds and Besner (2006) provided evidence on the central attention demands of reading. Earlier, we indicated that form-to-form mapping in reading involves orthographic processing and word-form encoding, including morphological, phonological, and phonetic encoding. Reynolds and Besner (2006) obtained evidence that word-form encoding in reading aloud requires central attentional capacity. They used the PRP procedure with participants performing manual tone discrimination (Task 1) and reading aloud (Task 2) tasks. Experiment 1 manipulated the duration of the form perception stage of word reading through long-lag repetition priming, which refers to shorter RTs for repeated than for novel words over lags greater than 100 intervening trials. According to Reynolds and Besner (2006), this type of priming affects orthographic–lexical processing, because it occurs for words but not for pseudowords and it is not affected by changes in case. Participants read aloud novel and repeated words presented 50 or 750 ms after tone onset. Reading RTs were shorter for the repeated than for the novel words, and this effect was present in the reading RTs at the long 750-ms SOA but not at the short 50-ms SOA. These results suggest that orthographic–lexical processing of the words (Task 2) occurred in parallel with tone processing (Task 1), before the lockout point of the word reading process, and the effect of repetition priming was absorbed by the pause, as we explain below.

Assume that participants strategically lock out the word reading process just before the onset of word-form encoding, so that processes in the tone task (Task 1) and processes up to (but not including) word-form encoding in reading (Task 2) are allowed to run in parallel. As a result of the repetition priming, word processing will reach the lockout point earlier for the repeated than the novel words. However, at the short 50-ms SOA, word reading will reach the lockout point before the tone processing has reached the unlocking event. Consequently, processing in the reading task has to wait for the unlocking event to occur and the difference in processing time for the repeated and novel words will be absorbed by the pause. In contrast, at the long 750-ms SOA, word reading will not have to wait for the tone processing, and the repetition priming effect will be observed in the reading RTs. Thus, overlap of orthographic–lexical processing and tone processing at the short SOA, but not at the long one, explains why the effects of repetition priming and SOA are underadditive.

In Experiments 2–4 of Reynolds and Besner (2006), pseudoword length and grapheme–phoneme complexity were manipulated. In dual-route models of reading, such as the one proposed by Coltheart et al. (2001), letter processing occurs in parallel across a letter string, but sublexical grapheme-to-phoneme translation occurs serially, from left to right across the string. Therefore, the RT for reading pseudowords aloud increases with the number of letters, as empirically observed in earlier research. Moreover, grapheme-to-phoneme translation is more complex and takes longer when at least one phoneme corresponds to a multiletter grapheme (e.g., TH in STETH) than when each phoneme corresponds to a single letter (e.g., STEK). Reynolds and Besner (2006) observed that the effects of pseudoword length and grapheme–phoneme complexity were additive with SOA, suggesting that participants did not divide central attention between tone discrimination and phonological encoding in reading aloud. Instead, phonological encoding was locked out, so that it did not occur in parallel with the tone discrimination task. Consequently, the effects of length and grapheme–phoneme complexity were additive with SOA. In Experiments 5–7, Reynolds and Besner (2006) examined whether participants divide attention between tone discrimination and lexical aspects of word-form encoding by manipulating orthographic neighborhood density, which refers to the number of words created by changing each letter of a word, one at a time. Reynolds and Besner (2006) reviewed evidence suggesting that the RT of reading aloud words and pseudowords decreases as neighborhood density increases. This effect of neighborhood density was argued to arise in word-form encoding. In the experiments of Reynolds and Besner (2006), the effect of neighborhood density was additive with SOA, suggesting that participants did not divide central attention between tone discrimination and lexical aspects of word-form encoding in reading aloud. To conclude, the results of Reynolds and Besner (2006) suggest that lexical and phonological stages of word-form encoding in reading aloud require central attention, whereas the orthographic–lexical processing of letter strings does not.

In all their experiments, Reynolds and Besner (2006) observed that the tone discrimination RTs (Task 1) were shorter at the 50-ms than the 750-ms SOA. If central attention is not divided between tasks, as the results of Reynolds and Besner (2006) suggest, then Task 1 RTs should be the same for long and short SOAs, because Task 1 receives full capacity in both cases. In contrast, Task 1 RTs were smaller at the short than the long SOA in the experiments of Reynolds and Besner. However, Task 1 RTs should only be constant across SOAs if attentional capacity is fixed, which does not need to hold (Tombu and Jolicoeur, 2003). Evidence suggests that the available capacity increases when participants put more effort into the tasks, which depends on the demands of concurrent activities (Kahneman, 1973). The demands are presumably higher at short than long SOAs. Exerting greater effort may decrease RTs at short SOAs, as was the case in the experiments of Reynolds and Besner (2006).

Whereas word reading requires central attention, it requires less attention than picture naming, according to the WEAVER++ model. This is because the pathway through the lexical network is shorter for reading than for picture naming, as illustrated in **Figure 1**. In line with the model, evidence from eye tracking suggests that shifts of gaze occur closer to articulation onset in naming pictures than in reading their names (Roelofs, 2007). An eye tracking study measured the mean latencies for the vocal responses and gaze shifts in picture naming and word reading in a semantic condition (e.g., a pictured dog combined with the word CAT), an unrelated condition (e.g., a pictured dog combined with the word PIN), and a control condition (e.g., a pictured dog combined with XXX for picture naming or the word DOG in an empty picture frame for word reading). A distractor effect was obtained in picture naming but not in word reading, suggesting differences in attention demands between the two tasks. In all three distractor conditions, the gaze shifts occurred about 66 ms before articulation onset in picture naming, whereas they happened already about 156 ms before articulation onset in word reading (Roelofs, 2007). Given the shorter network distance for word reading than picture naming (see **Figure 1**), attentional enhancements may be less for reading than naming. If enhancements are required until the word has been planned sufficiently, this explains why attention, as indexed by eye gazes, is sustained longer to word planning in picture naming than in word reading, regardless of whether or not distractors are present. However, such difference in gaze shift latencies was not observed when participants switched between naming the picture and reading the word aloud of picture–word combinations (Roelofs, 2008b). Pictures and words were presented in red and green. The task was picture naming or word reading depending on whether the picture or word was presented in green color, which varied randomly from trial to trial. In this task situation, gaze shifted around 100 ms before articulation onset in both picture naming and word reading. Apparently, there is a greater need to sustain attention to word reading when the distractor pictures have to be named on other trials and therefore are more likely to interfere with word reading.

EVIDENCE THAT WORD PLANNING DOES NOT REQUIRE FULL CENTRAL ATTENTION

WEAVER++ assumes that all word planning processes up to and including phonological encoding require some attentional capacity. However, the planning processes do not require full attentional capacity, meaning that central attention may be shared between word planning and other attention demanding concurrent processes. In contrast, other researchers (i.e., Ferreira and Pashler, 2002; Dell'Acqua et al., 2007; Ayora et al., 2011) proposed a central bottleneck model in which a process requires undivided attention or no attention, with no middle ground. For example, Ferreira and Pashler (2002) argued that lemma and morpheme selection in word planning preclude any other concurrent process that also requires central attention, such as response selection in a non-linguistic task.

Recent empirical results indicate that conflicts in word planning may be resolved while concurrently performing an unrelated non-linguistic task, making a task decision, or a go/no-go decision. These findings suggest that word planning does not require full central attention. A type of conflict that has been extensively studied is the increased response competition underlying the semantic interference effect, described above: RTs are longer for picture naming when the word is semantically related to the picture name (e.g., picture of a dog combined with the word CAT) relative to unrelated words (e.g., the word PIN). Whereas in single-task performance, distractor words in picture naming yield semantic interference, this effect may be absent when simultaneously performing picture naming and a concurrent task or process.

CENTRAL ATTENTION SHARING IN DUAL-TASK PERFORMANCE

Dell'Acqua et al. (2007) observed that the semantic interference effect in picture naming may diminish or disappear at short SOAs in the PRP procedure. Participants performed a manual tone discrimination task (Task 1) and a picture–word interference task (Task 2). The tones preceded the picture–word stimuli by SOAs of 100, 350, or 1000 ms. The semantic interference effect was much smaller at the 350-ms SOA than at the 1000-ms SOA, and the interference was absent at the 100-ms SOA. These results suggest that the semantic interference in picture naming was resolved while simultaneously performing the tone discrimination task.

This evidence suggests that central attention may be divided between tone discrimination, on the one hand, and resolving the conflict underlying the semantic interference effect in picture naming, on the other hand. **Figure 5** illustrates our account of the data of Dell'Acqua et al. (2007), which are shown in **Figure 6** together with the WEAVER++ simulation results obtained by Piai et al. (2011). At the short SOA, picture naming has to pause after resolving the conflict in lemma selection. Consequently, the semantic interference in picture naming will be absorbed by the pause. In contrast, at the long SOA, attention will have shifted away from the tone task before the picture–word stimulus is presented. As a result, the conflict in lemma selection cannot be resolved while performing the tone task and semantic interference will be reflected in the naming RTs.

A hallmark of attentional capacity sharing is that Task 1 RT increases as SOA decreases in dual-task performance. If some proportion of the attentional capacity is allocated to Task 1 and the remainder to Task 2 when both tasks require central attention, this will increase Task 1 response latencies at short SOAs compared to long ones (when 100% of the capacity may be allocated to Task 1). We assumed that participants in the experiment of Dell'Acqua et al. (2007) shared attentional capacity between the tone discrimination task (Task 1) and the picture naming task (Task 2). However, in that study, Task 1 RTs did not increase at short SOAs, which seems to challenge the assumption that capacity was shared.

However, the Task 1 RTs should only be increased at short SOAs if attentional capacity is fixed and the capacity allocated to Task 1 and Task 2 sums to full capacity (cf. Tombu and Jolicoeur, 2003), which does not need to hold. As we indicated earlier, evidence suggests that the available capacity increases when participants put more effort into tasks (Kahneman, 1973). Exerting greater effort may compensate for the slowing of tasks caused by dividing attentional capacity at short SOAs. If the participants of Dell'Acqua et al. (2007) increased capacity by exerting greater effort at short SOAs, the Task 1 RTs do not need to become longer, as empirically observed. According to Kahneman (1973), the amount of

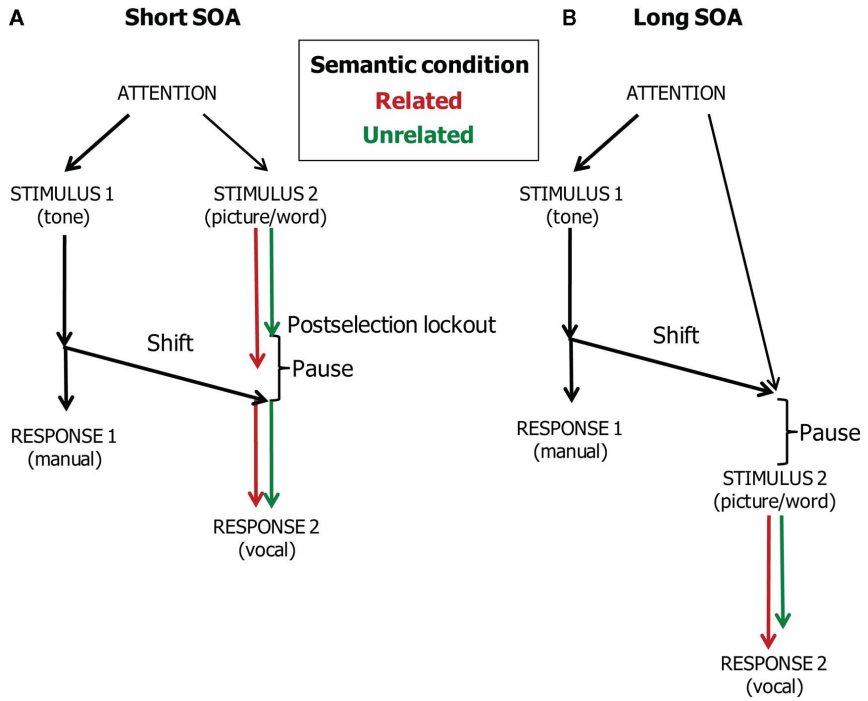


FIGURE 5 | Theoretical account of the time course of manual responding (Task 1) and vocal responding (Task 2) in the dual-task study of Dell'Acqua et al. (2007). Pictures are named in semantically related and unrelated conditions. The stimulus onset asynchrony (SOA) is short **(A)** or long **(B)**.

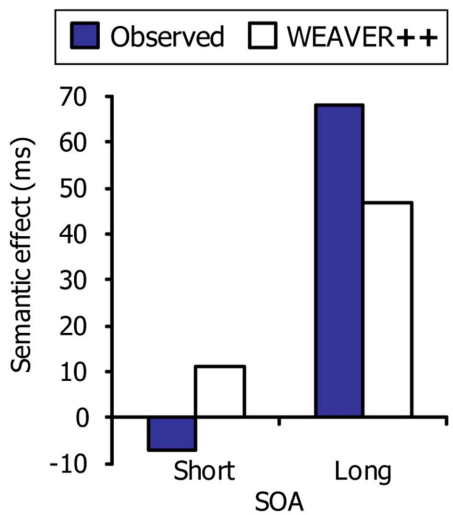


FIGURE 6 | The semantic interference effect in picture naming as a function of stimulus onset asynchrony (SOA) in the psychological refractory period procedure. Shown are the real data (Dell'Acqua et al., 2007) and WEAVER++ simulation results. The short SOA was 100 ms and the long SOA was 1000 ms.

attentional capacity available at any time depends on the demands of current activities, which is presumably less at long than short SOAs. To conclude, given the potentially confounding effect of effort across SOAs in the study of Dell'Acqua et al. (2007), the absence of an increase of Task 1 RTs at short SOAs does not exclude that attentional capacity was shared.

CENTRAL ATTENTION SHARING IN MAKING TASK-CHOICE AND GO/NO-GO DECISIONS

In line with our account of the findings of Dell'Acqua et al. (2007) illustrated in **Figure 5**, it was found that the semantic interference effect in picture naming may also disappear when simultaneously making a task choice (Piai et al., 2011). In the task choice procedure (Besner and Care, 2003), participants receive a cue at every trial indicating which task to perform. This cue can either be given before the target or simultaneously with it. In this procedure, only the response to the target stimulus is required, so no response selection takes place for the cue stimulus. The logic of the task-choice paradigm is similar to the dual-task interference logic (Besner and Care, 2003). Under our account, one or more stages of processing for the target stimulus are postponed until the decision concerning what task to perform has been made. If processes involved in the task to be performed (e.g., picture naming) are run in parallel with the task-choice process, effects related to these processes, such as semantic interference, may (partly) be absorbed.

In the picture–word interference study of Piai et al. (2011), participants had to decide between naming the picture or reading the word aloud depending on the presentation color of the word. Whereas semantic interference was obtained in a standard picture–word interference experiment, the semantic interference effect disappeared when task choices had to be made. Assuming that semantic interference arises at the level of response selection, these findings suggest that participants locked out picture naming

processes after response selection and that the semantic interference effect was absorbed by the pause created by the task-choice process. **Figure 7** depicts the account.

Figure 8 shows the empirical data of Piai et al. (2011) together with the WEAVER++ simulation results. Without task decision, a full-blown semantic interference effect occurs in the model, as typically observed with picture naming in picture–word interference experiments. However, when a task choice has to be made, the pause created by the task-choice process may absorb the semantic interference effect in the model, as empirically observed.

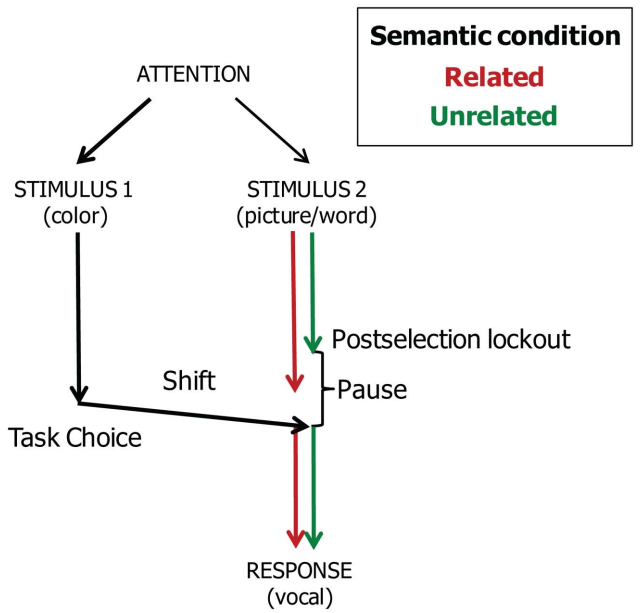


FIGURE 7 | Theoretical account of the timing of task-choice processes (Task 1) and vocal responding (Task 2) in the task-choice study of Piai et al. (2011). Pictures are named in semantically related and unrelated conditions.

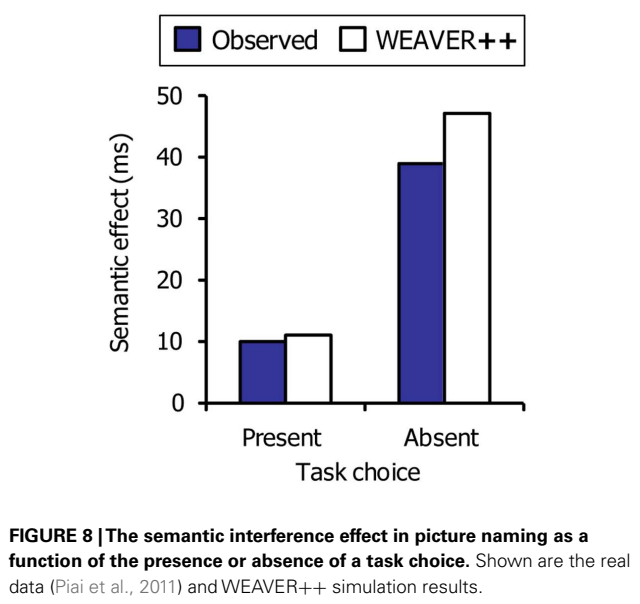


FIGURE 8 | The semantic interference effect in picture naming as a function of the presence or absence of a task choice. Shown are the real data (Piai et al., 2011) and WEAVER++ simulation results.

Importantly, under the assumption of a postselection lockout point for the picture naming task, semantic interference will only be absorbed if the choice processes take longer than the duration of processes up to and including lemma selection for picture naming in the semantically related condition, as illustrated in **Figure 7**. In contrast, if choice processes take less time than the processes up to and including lemma selection, semantic interference should be obtained. This corresponds to what Janssen et al. (2008) observed using the task-choice procedure and to what Mädebach et al. (2011) observed when the choice processes consisted of a go/no-go decision based on the color of the word. In the WEAVER++ model, decreasing the duration of the choice process a little (e.g., by 25 ms) yields a semantic interference effect (e.g., of some 30 ms), as observed in these studies.

Evidence that attention shifts occur earlier in go/no-go than dual-task situations was obtained in an eye tracking study of Lamers and Roelofs (2011). Participants vocally responded to congruent and incongruent flanker stimuli presented on the left side of a computer screen and shifted gaze to left- or right-pointing arrows presented on the right side of the screen. The arrows required a manual response (dual task) or determined whether the naming response to the flanker stimuli had to be given or not (go/no-go). The results showed that the naming RTs and gaze shift latencies were longer on incongruent than congruent trials in both dual-task and go/no-go performance. In dual-task performance, the flanker effect was also present in the manual RTs for the arrow stimulus, reflecting a propagation of the distractor effect from the naming to the manual responses. These results suggest that gaze shifts occur after response selection in both dual-task and go/no-go performance with vocal responding. However, the gaze shift latencies were on average 185 ms shorter in the go/no-go condition than in the dual-task condition. Thus, although gazes shifted after response selection in both the go/no-go and the dual-task conditions (as suggested by the presence of the flanker effects in the gaze shift latencies), attention seemed to shift earlier in the go/no-go than the dual-task condition.

ATTENTION IN IMPAIRED LANGUAGE PERFORMANCE

Whereas attentional capacity may increase with effort, there is an upper limit (Kahneman, 1973). Moreover, the increase may often be insufficient to fully meet the demands of a task, especially when the task is difficult. A task may be difficult, for example, when it is complex or when the task is simple but the individual performing the task has a deficit in one or more of the component abilities that are required. For example, evidence suggests that individuals with developmental dyslexia have difficulty in performing grapheme-to-phoneme translations in reading, presumably because they fail to develop strong connections. Evidence suggests that dyslexic individuals try to compensate the weaker connections by allocation of more attention to the grapheme–phoneme translation process. Brain areas associated with word-form perception, such as the left occipito-temporal sulcus, are less activated in dyslexic than normal readers. In contrast, brain areas associated with attentional control, such as regions in prefrontal and parietal cortex, are more highly activated in dyslexic than normal

readers in reading performance (see Shaywitz and Shaywitz, 2008, for a review). This suggests that dyslexic readers try to overcome or diminish their reading problem by investing more attention. However, given that problems remain (e.g., reading RTs are longer for dyslexic than normal readers), the increased attention appears insufficient to counteract the slowing caused by weak grapheme–phoneme connections. Similarly, increased attention and effort is typically insufficient to compensate for the detrimental consequences of brain damage in acquired dyslexia and aphasia (e.g., Murray, 1999). Attention problems may worsen performance in dyslexia and aphasia (e.g., Murray, 1999; Shaywitz and Shaywitz, 2008).

Evidence suggests that attention deficits also contribute to the impaired language performance of individuals with SLI. This is a disorder of language acquisition and use in children who otherwise appear to be normally developing. The disorder may persist into adulthood. The features of the impaired language performance in SLI are quite variable, but common characteristics are a delay in starting to talk in childhood, deviant production of speech sounds, a restricted vocabulary, slow and inaccurate picture naming, and use of simplified grammatical structures, including omission of articles and plural and past tense endings (see Leonard, 1998, for a review). In general, individuals with SLI seem to have a problem in dealing with (relatively) complex language structures, in both speech production and comprehension. A prominent account of SLI holds that these difficulties with complexity in language reflect a reduced capacity of systems underlying language processes, resulting from a limitation in general processing capacity (Leonard, 1998). It is becoming increasingly clear that (subclinical) attention deficits also contribute to SLI.

Individuals with SLI appear to have reduced working memory capacity, as assessed by pseudoword repetition and listening span tasks (e.g., Ellis Weismer et al., 2005, for a review). Moreover, evidence suggests that children with SLI have deficits in sustained attention (e.g., Spaulding et al., 2008; Finneran et al., 2009). The reduced working memory and sustained attention capacities may have a common ground. In an influential functional analysis of executive control by Miyake et al. (2000), three types of executive abilities are distinguished: monitoring and updating of working memory representations, inhibiting of dominant responses, and shifting of tasks or mental sets. Evidence suggests that working memory performance is specifically related to the updating ability (Miyake et al., 2000), whereas sustained attention performance is related to the updating and inhibiting abilities (Unsworth et al., 2010). Im-Bolter et al. (2006) provided evidence that the updating and inhibiting abilities are deficient in SLI.

Working memory and sustained attention play an important role in WEAVER++. In this model, the lexical network is accessed by spreading activation while the condition–action rules determine what is done with the activated lexical information depending on the task goal in working memory. The task goal is achieved by successively updating subgoals in the course of the word planning process. In conceptually driven word planning, an initial subgoal is to select a lemma for a selected concept. The next subgoal is to select one or more morphemes for the selected lemma. Next, the subgoal is to select phonemes for the selected morphemes. Then, the subgoal is to syllabify the selected phonemes and to assign word accent. A final subgoal is to select syllable motor programs for the syllabified phonemes. For the planning process to be successful, attention needs to be sustained until the phonological form has been planned and syllable motor programs may be accessed. As discussed by Leonard (1998) for a WEAVER++ type of model, difficulties in word planning may arise when there are capacity restrictions in the language processes involved. For example, a capacity restriction in activating or selecting morphemes for the selected lemma may result in an omission of inflectional morphemes, such as past tense endings. This type of problem will be reinforced by capacity restrictions in working memory and sustained attention (i.e., the updating ability). For example, problems in successively maintaining subgoals will impede the planning process, especially when a subgoal concerns a complex mapping between levels (e.g., such as the mapping between lemmas and morphemes, e.g., Janssen et al., 2002, 2004).

A role of attention in dyslexia, aphasia, and SLI has practical implications. To the extent that attention deficits contribute to the impaired language performance, therapeutic interventions that only deal with the underlying language processes are not providing the afflicted individuals with what they need. Rather, interventions should aim at improving the attention abilities as well (e.g., Murray, 1999; Shaywitz and Shaywitz, 2008; Finneran et al., 2009).

CONCLUSION

Evidence suggests that word planning requires some but not full central attention. Empirical results indicate that processes up to and including phonological encoding in word planning delay, or are delayed by, the performance of concurrent unrelated non-linguistic tasks. These findings suggest that word planning requires some attentional capacity. Moreover, empirical results indicate that conflicts in word planning may be resolved while concurrently performing an unrelated non-linguistic task, making a task decision, or making a go/no-go decision. These findings suggest that word planning does not require full attentional capacity.

ACKNOWLEDGMENTS

Preparation of this article was supported by a grant (Open Competition MaGW 400-09-138) from the Netherlands Organisation for Scientific Research.

REFERENCES

Ayora, P., Peressotti, F., Alario, F.-X., Mulatti, C., Pluchino, P., Job, R., and Dell'Acqua, R. (2011). What phonological facilitation tells about semantic interference: a dual-task study. *Front. Psychol.* 2:57. doi:10.3389/fpsyg.2011.00057

Baddeley, A. (1996). Exploring the central executive. *Q. J. Exp. Psychol.* 49A, 5–28.

Badre, D., Poldrack, R. A., Paré-Blagoev, E., Insler, R. Z., and Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron* 47, 907–918.

Besner, D., and Care, S. (2003). A paradigm for exploring what the mind does while deciding what it should do. *Can. J. Exp. Psychol.* 57, 311–320.

Besner, D., and Stolz, J. A. (1999). What kind of attention modulates the Stroop effect? *Psychon. Bull. Rev.* 6, 99–104.

Besner, D., Stolz, J. A., and Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychon. Bull. Rev.* 4, 221–225.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D.

(2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256.

Cook, A. E., and Meyer, A. S. (2008). Capacity demands of phoneme selection in word production: new evidence from dual-task experiments. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 886–899.

de Zubicaray, G. I., Wilson, S. J., McMahon, K. K., and Muthiah, S. (2001). The semantic interference effect in the picture-word paradigm: an event-related fMRI study employing overt responses. *Hum. Brain Mapp.* 14, 218–227.

Dehaene, S. (2009). *Reading in the Brain.* New York: Viking.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., and Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychol. Rev.* 104, 801–838.

Dell'Acqua, R., Job, R., Peressotti, F., and Pascali, A. (2007). The picture-word interference effect is not a Stroop effect. *Psychon. Bull. Rev.* 14, 717–722.

Ellis Weismer, S., Plante, E., Jones, M., and Tomblin, J. B. (2005). A functional magnetic resonance imaging investigation of verbal working memory in adolescents with specific language impairment. *J. Speech Lang. Hear. Res.* 48, 405–425.

Ferreira, V., and Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 1187–1199.

Finneran, D. A., Francis, A. L., and Leonard, L. B. (2009). Sustained attention in children with specific language impairment (SLI). *J. Speech Lang. Hear. Res.* 52, 915–929.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition* 82, B1–B14.

Griffin, Z. M. (2004). "Why look? Reasons for eye movements related to language production," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. M. Henderson and F. Ferreira (Hove: Psychology Press), 213–247.

Im-Bolter, N., Johnson, J., and Pascual-Leone, J. (2006). Processing limitations in children with specific language impairment: the role of executive function. *Child Dev.* 77, 1822–1841.

Indefrey, P., and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition* 92, 101–144.

Janssen, D. P., Roelofs, A., and Levelt, W. J. M. (2002). Inflectional frames in language production. *Lang. Cogn. Process.* 17, 209–236.

Janssen, D. P., Roelofs, A., and Levelt, W. J. M. (2004). Stem complexity and inflectional encoding in language production. *J. Psycholinguist. Res.* 33, 365–381.

Janssen, N., Schirm, W., Mahon, B. Z., and Caramazza, A. (2008). Semantic interference in a delayed naming task: evidence for the response exclusion hypothesis. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 249–256.

Johnston, J. C., McCann, R. S., and Remington, R. W. (1995). Chronometric evidence for two types of attention. *Psychol. Sci.* 6, 365–369.

Kahneman, D. (1973). *Attention and Effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kan, I. P., and Thompson-Schill, S. L. (2004). Effect of name agreement on prefrontal activity during overt and covert picture naming. *Cogn. Affect. Behav. Neurosci.* 4, 43–57.

Kane, M. J., and Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective. *Psychon. Bull. Rev.* 9, 637–671.

Korvorst, M., Roelofs, A., and Levelt, W. J. M. (2006). Incrementality in naming and reading complex numerals: evidence from eye tracking. *Q. J. Exp. Psychol.* 59, 296–311.

La Heij, W., van der Heijden, A. H. C., and Plooij, P. (2001). A paradoxical exposure-duration effect in the Stroop task: temporal segregation between stimulus attributes facilitates selection. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 622–632.

Lamers, M., and Roelofs, A. (2007). Role of Gestalt grouping in selective attention: evidence from the Stroop task. *Percept. Psychophys.* 69, 1305–1314.

Lamers, M., and Roelofs, A. (2011). Attention and gaze shifting in dual-task and go/no-go performance with vocal responding. *Acta Psychol. (Amst.)* 137, 261–268.

Leonard, L. B. (1998). *Children With Specific Language Impairment.* Cambridge, MA: MIT Press.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci.* 22, 1–38.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychol. Bull.* 109, 163–203.

Mädebach, A., Oppermann, F., Hantsch, A., Curda, C., and Jescheniak, J. D. (2011). Is there semantic interference in delayed naming? *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 522–538.

Malpass, D., and Meyer, A. S. (2010). The time course of name retrieval during multiple-object naming: evidence from extrafoveal-on-foveal effects. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 523–537.

Maril, A., Wagner, A. D., and Schacter, D. L. (2001). On the tip of the tongue: an event-related fMRI study of semantic retrieval failure and cognitive conflict. *Neuron* 31, 653–660.

Meyer, A. S., Roelofs, A., and Levelt, W. J. M. (2003). Word length effects in object naming: the role of a response criterion. *J. Mem. Lang.* 48, 131–147.

Meyer, A. S., Sleiderink, A. M., and Levelt, W. J. M. (1998). Viewing and naming objects. *Cognition* 66, B25–B33.

Meyer, D. E., and Kieras, D. E. (1997a). A computational theory of executive cognitive processes and multiple-task performance: part 1. Basic mechanisms. *Psychol. Rev.* 104, 3–65.

Meyer, D. E., and Kieras, D. E. (1997b). A computational theory of executive cognitive processes and multiple-task performance: part 2. Accounts of psychological refractory-period phenomena. *Psychol. Rev.* 104, 749–791.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100.

Murray, L. L. (1999). Attention and aphasia: theory, research, and clinical implications. *Aphasiology* 13, 91–111.

Naccache, L., Dehaene, S., Cohen, L., Habert, M.-O., Guichart-Gomez, E., Galanaud, D., and Willer, J.-C. (2005). Effortless control: executive attention and conscious feeling of mental effort are dissociable. *Neuropsychologia* 43, 1318–1328.

Pashler, H. (1998). *The Psychology of Attention.* Cambridge, MA: MIT Press.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* 331, 585–589.

Petersen, S. E., van Mier, H., Fiez, J. A., and Raichle, M. E. (1998). The effects of practice on the functional anatomy of task performance. *Proc. Natl. Acad. Sci. U.S.A.* 95, 853–860.

Piai, V., Roelofs, A., and Schriefers, H. (2011). Semantic interference in immediate and delayed naming and reading: attention and task decisions. *J. Mem. Lang.* 64, 404–423.

Posner, M. I., and Raichle, M. E. (1994). *Images of Mind.* New York: W. H. Freeman.

Posner, M. I., and Rothbart, M. K. (2007). *Educating the Human Brain.* Washington, DC: APA Books.

Reynolds, M., and Besner, D. (2006). Reading aloud is not automatic: processing capacity is required to generate a phonological code from print. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 1303–1323.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition* 42, 107–142.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition* 64, 249–284.

Roelofs, A. (2003). Goal-referenced selection of verbal action: modeling attentional control in the Stroop task. *Psychol. Rev.* 110, 88–125.

Roelofs, A. (2004). Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: comment on Rapp and Goldrick (2000). *Psychol. Rev.* 111, 561–572.

Roelofs, A. (2006). Context effects of pictures and words in naming objects, reading words, and generating simple phrases. *Q. J. Exp. Psychol.* 59, 1764–1784.

Roelofs, A. (2007). Attention and gaze control in picture naming, word reading, and word categorizing. *J. Mem. Lang.* 5, 232–251.

Roelofs, A. (2008a). Attention, gaze shifting, and dual-task interference from phonological encoding in spoken word planning. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1580–1598.

Roelofs, A. (2008b). Tracing attention and the activation flow in spoken word planning using eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 353–368.

Roelofs, A. (2008c). Dynamics of the attentional control of word retrieval: analyses of response time distributions. *J. Exp. Psychol. Gen.* 137, 303–323.

Roelofs, A. (2011). Attention, exposure duration, and gaze shifting in naming performance. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 860–873.

Roelofs, A., and Hagoort, P. (2002). Control of language use: cognitive modeling of the hemodynamics of Stroop task performance. *Brain Res. Cogn. Brain Res.* 15, 85–97.

Roelofs, A., and Lamers, M. (2007). "Modelling the control of visual attention in Stroop-like tasks," in *Automaticity and Control in Language Processing*, eds A. S. Meyer, L. R. Wheeldon, and A. Krott (Hove: Psychology Press), 123–142.

Roelofs, A., Özdemir, R., and Levelt, W. J. M. (2007). Influences of spoken word planning on speech recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 900–913.

Roelofs, A., van Turennout, M., and Coles, M. G. H. (2006). Anterior cingulate cortex activity can be independent of response conflict in Stroop-like tasks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13884–13889.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., and Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science* 326, 445–449.

Sakai, K. (2008). Task set and prefrontal cortex. *Annu. Rev. Neurosci.* 31, 219–245.

Shaywitz, S. E., and Shaywitz, B. A. (2008). Paying attention to reading: the neurobiology of reading and dyslexia. *Dev. Psychopathol.* 20, 1329–1349.

Spaulding, T. J., Plante, E., and Vance, R. (2008). Sustained selective attention skills of preschool children with specific language impairment: evidence for separate attentional capacities. *J. Speech Lang. Hear. Res.* 51, 16–34.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., and Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proc. Natl. Acad. Sci. U.S.A.* 94, 14792–14797.

Tombu, M., and Jolicoeur, P. (2003). A central capacity sharing model of dual-task performance. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 3–18.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231–270.

Unsworth, N., Redick, T. S., Lakey, C. E., and Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: an individual differences investigation. *Intelligence* 38, 111–122.

Wernicke, C. (1874). *Der aphasische Symptomencomplex* [The Aphasic Symptom Complex]. Breslau: Cohn and Weigert.

Wright, R. D., and Ward, L. M. (2008). *Orienting of Attention*. Oxford: Oxford University Press.

Wundt, W. (1900). *Die Sprache* [Language]. Leipzig: Verlag von Wilhelm Engelmann.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 July 2011; paper pending published: 10 August 2011; accepted: 13 October 2011; published online: 07 November 2011.

Citation: Roelofs A and Piai V (2011) Attention demands of spoken word planning: a review. Front. Psychology ***2****:307. doi: 10.3389/fpsyg.2011.00307*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 18 January 2012
doi: 10.3389/fpsyg.2011.00396

Referential and visual cues to structural choice in visually situated sentence production

Andriy Myachykov*, Dominic Thompson, Simon Garrod and Christoph Scheepers

Department of Psychology, Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK

Edited by:
Yury Y. Shtyrov, Medical Research Council, UK

Reviewed by:
Lucy Jane MacGregor, Medical Research Council, UK
Andrej Kibrik, Moscow State University, Russia

****Correspondence:***
Andriy Myachykov, Department of Psychology, Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, UK.
e-mail: andriy.myachykov@glasgow.ac.uk

We investigated how conceptually informative (referent preview) and conceptually uninformative (pointer to referent's location) visual cues affect structural choice during production of English transitive sentences. Cueing the Agent or the Patient prior to presenting the target-event reliably predicted the likelihood of selecting this referent as the sentential Subject, triggering, correspondingly, the choice between active and passive voice. Importantly, there was no difference in the magnitude of the general Cueing effect between the informative and uninformative cueing conditions, suggesting that attentionally driven structural selection relies on a direct automatic mapping mechanism from attentional focus to the Subject's position in a sentence. This mechanism is, therefore, independent of accessing conceptual, and possibly lexical, information about the cued referent provided by referent preview.

Keywords: sentence production, visual attention, structural choice

INTRODUCTION

Many psycholinguistic theories of sentence production suggest that selecting words, grammatical roles, and structural configurations are not arbitrary processes as they necessarily reflect the organization of the conveyed conceptual message via the rules of a regular interface between language and cognition (e.g., Bock, 1982; Jackendoff, 2002; Vigliocco and Hartsuiker, 2002; Myachykov et al., 2007). The emphasis of this paper is on the interface between the speaker's visual attention on the event's referents, accessibility of the conceptual information associated with these referents, and the assignment of grammatical roles and consequent syntactic structures in a spoken sentence.

Speaking about events in a real time situated context is a seemingly effortless routine task, performed daily by every language user. Yet, producing even a single utterance about a simple event is a complex process involving rapid and well-orchestrated execution of both linguistic and non-linguistic operations in the speaker's mind (Jackendoff, 2002). These operations do not only include information retrieval, they are also inherently selective; that is, they involve selecting information for earlier or later processing. Consider a situation in which the speaker describes a simple event, for example, *a boy kicking a ball*. The first necessary step in generating a sentence about this event is creating a non-linguistic conceptual plan of the event, or its *message* (Levelt, 1989). This message will be eventually translated into an emerging sentence via selecting words and assigning to them specific grammatical roles and positions in a syntactic structure. The speaker's visual attention will guide the translation by progressively selecting information for processing. This selection will be based on a number of parameters that make a particular referent, word, or structure more relevant, available, or conspicuous than the other available alternatives. This selection process already starts at the earliest stages of message apprehension when the non-linguistic properties of the event (including the relative *salience* of the interacting referents) are encoded. At this stage, a variety of factors act as *cues* increasing referential salience. Some cues may be part of the speaker's own perspective on the event or knowledge about the referents. These are *endogenous* cues. Other cues are *exogenous*; they are specific features of the referent itself, for example, its size, shape, motion, or color. Let us assume the *boy's* larger size acts as an exogenous cue, preferentially attracting the speaker's attentional focus to it over the smaller and less salient *ball*. As a result, the *boy* may be selected for earlier and deeper processing than the *ball* (e.g., Itti et al., 1998; Itti and Koch, 2000; Parkhurst et al., 2002)[1]. In other words, the *boy* will be coded in the message as the referent that is more *accessible* for processing than the *ball* (Bock and Warren, 1985).

As the non-linguistic message is forwarded for linguistic formulation, the more accessible referent may receive preferential treatment by means of earlier *lemma retrieval* and also by being assigned a more important grammatical role during *structural assembly* (Levelt, 1989). Hence, at lemma retrieval (where concepts receive their lexical names accompanied by grammatical properties), the *boy's* name will be accessed earlier than that of the *ball*. At the stage of structural assembly, the *boy* may receive a more prominent role, e.g., the Subject. In English, this will almost inevitably lead to the selection of the active-voice frame *A boy kicked a ball* (where the *agent* assumes the Subject role), rather than the alternative passive-voice frame *A ball was kicked by a boy* (where the *patient* assumes the Subject role). This simple example portrays how attentional focus driven by purely perceptual properties of a referent may in principle predict the likelihood of Subject assignment and the resulting choice between available

[1]We acknowledge that factors other than exogenous cues play a role in capturing visual focus during natural scene viewing (e.g., Corbetta and Shulman, 2002; Henderson, 2003). Here, we focus on the role of referential salience and, therefore, on exogenously captured visual attention in the process of sentence generation.

structural configurations (see Myachykov et al., 2011, for a recent review).

It has to be noted that visual salience is not the only factor that can influence Subject assignment. It is well known that *linguistic* cues, such as priming a word associated with a referent (Flores d'Arcais, 1975; Osgood and Bock, 1977; Bock and Irwin, 1980; Bates and Devescovi, 1989; Prat-Sala and Branigan, 2000), or priming aspects of structural configuration (Ferreira and Bock, 2006; Branigan, 2007; Pickering and Ferreira, 2008 for recent reviews), also exert strong influences on Subject assignment and the resulting structural choice. One can therefore hypothesize that the Subject role encodes both the non-linguistic (perceptual or conceptual) and the linguistic (lexical or structural) salience of a referent. Here, we focus on the role of non-linguistic salience as determined by visual and/or conceptual cues.

The tendency of salient referents to assume prominent grammatical roles in sentences was already noted in a number of early psycholinguistic experiments using a *referent preview* paradigm. One such study (Prentice, 1967) used a set of cartoon pictures portraying simple transitive interactions between two characters (e.g., *fireman kicking cat*). Some of the characters were human, others animals, and inanimate objects. The pictures were paired with slides of one of the event's characters: the agent or the patient. Therefore, one of the referents was cued before the full event was displayed. Participants first viewed the cue slide and then the whole event, of which they provided spoken descriptions. As a result, speakers were more likely to place the previewed referent first in their target-event descriptions, making it the sentential Subject, leading to a higher proportion of passive-voice descriptions (e.g., *A cat was kicked by a fireman*) in the patient-preview condition. Prentice explained this result by suggesting that referent preview acted as an attentional cue to the referent that participated in the subsequent event. Importantly, the cue slide was always presented *in the center* of the screen and not in the location where the corresponding referent would later appear. Hence, visual attention *per se* does not have to be invoked, as the structural choice effect most likely resulted from preferential access to the conceptual (and potentially, lexical) information about the cued referent, rather than from directing attention to the subsequent target's location. We will return to this issue below.

Experiments that followed Prentice (1967) used a similar setup. For example, Turner and Rommetveit (1968) presented children with active/passive sentences and later asked participants to recall these sentences. Both at the time of encoding and recall, sentences were presented to participants randomly paired with a picture of the agent, the patient, the whole event, or a blank. Among other things, Turner and Rommetveit found that the active-voice sentences were more likely to be recalled correctly if the visually primed referent was the agent, while the passive-voice sentences were better remembered if the primed referent was the patient. Although the latter study involved referent preview at both the encoding and the recall stages, the retrieval-picture effect and the storage-picture effect were attested separately. The authors found that the retrieval-picture effect was stronger, suggesting that the assignment of the referent's role in the sentence was affected more strongly by referent preview during production of the target description than by the encoding of the target sentence for later recall.

These early studies seem to confirm the hypothetical scenario we outlined above: preferential attention to a referent can predict the choice of sentential structure via assignment of the Subject role to the most salient referent. However, the "attentional" manipulations in these studies employed a referent preview long enough (more than 600 ms) not only to bias attention toward the subsequently presented referent, but also for the participant to fully recognize the referent's identity, and potentially even activate its name. Also, the preview of a referent did not inform the participants about the corresponding referent's location in the subsequently presented target event. Therefore, although visual attention may have been implicated in the resulting structural choice effect, a plausible alternative explanation might be that referent preview primed the speaker's access to the conceptual (and possibly lexical) information associated with the primed referent, which in itself is enough to predict Subject selection without invoking any specific notion of attention.

Studies using a *visual cueing* paradigm directly address the question of how visual attention *per se* can predict Subject assignment and structural choice. In contrast to a referent preview paradigm, visual cueing studies use visual prompts that do not provide any information about the cued referent (Posner, 1980). Participants usually see a pointer, a dot, or a square, cueing the referent's location before the event presentation or simultaneously with it. Importantly, the cue itself does not provide any conceptual information about the cued referent; hence, any resulting structural choice effect must be attributed to visual attention and not to other factors, for example, prior higher memorial activation of conceptual and/or lexical information associated with the cued referent.

One of the earliest studies using a visual cueing paradigm was the Fish Film experiment by Tomlin (1995). In this study, English speakers described an animated film portraying one fish eating another. A visual cue (a pointer) directed participants' attention to the eventual Patient or Agent fish as the two fish approached each other; that is, before the eating event. When the cue was on the eventual agent, participants predominantly described the event with an active-voice sentence (e.g., *The blue fish ate the red fish*). When it was on the patient, they produced passive-voice descriptions most of the time (e.g., *The red fish was eaten by the blue fish*). Hence, the focally attended referent was consistently assigned the sentential Subject role, driving the choice between active and passive voice. Although Tomlin's results were very intriguing, both the cueing procedure and the repetitive nature of the Fish Film paradigm received criticism from some psycholinguists for being "too brutal" (Bock et al., 2004) or crude and suggestive about the experimenter's goal (Gleitman et al., 2007). From a methodological point of view, such criticisms are justified to some extent. First, although the experimental instructions did not tell participants anything about how to treat the cue in relation to the choice of event description, it considerably constrained their attentional focus to the cued referent making it not only perceptually, but also conceptually, more accessible. In this respect, presenting a pointer cue together with the stimulus (for a time long enough to recognize the cued referent) makes this cueing manipulation

very similar to the referential priming paradigm described above. Hence, any conclusion about independent contributions of visual attention to the selection of the sentential Subject remains only partially justified. Second, the Fish Film paradigm instructs participants to view and describe continuously *all* the interactions between the fish, including those preceding the target event. This inevitably increases the *givenness* (e.g., Bock, 1982; Givon, 1992) of the cued fish. Finally, the repetitive nature of the target event and the lack of interrupting filler materials make effects of syntactic persistence a possible concern (Bock, 1986). Nevertheless, this original finding and the Fish Film paradigm became in many ways ground-breaking; its variants were later used in studies of other syntactic structures (e.g., Forrest, 1996) and languages structurally different from English (Diderichsen, 2001; Rasolofo, 2006; Myachykov and Tomlin, 2008).

A more recent study (Gleitman et al., 2007) tried to avoid the methodological problems in Tomlin (1995) by separating the cue from the target event, using *implicit* rather than *explicit* cues, and monitoring attention through eye-tracking. Sentences with verbs of perspective (*give/receive*), conjoined noun phrases (*The boy and the girl/The girl and the boy*), voice alternating transitive sentences, and symmetrical predicates (*The boy meets the girl/The girl meets the boy*) were elicited with the help of still pictures presented on a computer screen. Participants' attention was directed to the location of one of the subsequently presented referents, *before* the target-event presentation, by flashing a black square on the screen for 75 ms. This short cue duration ensured that participants remained unaware of the manipulation itself, although their gaze (and the focus of attention) was attracted to the cued location *implicitly*. The success of the cueing manipulation was monitored by recording eye movements in real time. Once the picture was on the screen, participants extemporaneously described the presented event without any further manipulations of attention. The magnitude of the resulting visual cueing effect was smaller than the one reported by Tomlin; nevertheless, the cued referent was more likely to be assigned the sentential Subject position, triggering the choice between corresponding structural alternatives.

Overall, the studies reviewed here, as well as a number of similar studies (see Myachykov et al., 2011 for a review) consistently showed that a visual cue to a specific referent in an event, uninformative with regard to the cued referent's conceptual and/or linguistic properties, reliably predicts the selection of that referent as the Subject (and associated structural choice). As a result, some theoretical proposals claim a direct link between visual attention on (or salience of) a referent on the one hand and assignment of the Subject role to that referent on the other (Tomlin, 1997; Myachykov et al., 2011). While this is a relatively simple and straightforward proposal, its validity is difficult to assess in the absence of studies that directly compare the effects of referential and visual cueing. One possibility is that referent preview provides more information about the cued referent than visual cueing. At least in principle, given enough preview time, speakers can extract both conceptual and lexical information about the referent. This is not so in the case of a purely visual cueing scenario. Indeed, if directing attention to the location of a referent (via a conceptually uninformative cue) provides only a part of the information provided by referent preview, then visual cueing might have a weaker effect on subsequent Subject selection than referent preview.

The issue of cue informativity introduced above is related to the psycholinguistic concept of *conceptual accessibility* or the ease of retrieval of the conceptual information about the referent from working memory (Bock and Warren, 1985). Although the concept itself is very broadly defined as related to notions such as "codeability," "imageability," "retrievability," etc., the concept itself has been repeatedly invoked in psycholinguistic studies in order to explain why information associated with some referents (or, more broadly, concepts) is accessed or retrieved ahead of the information about other referents or concepts. A number of referent-related properties were shown to be responsible for an increase in conceptual accessibility, such as givenness (Bock, 1977; Arnold et al., 2000), animacy (Clark, 1966; Sridhar, 1988; Bock et al., 1992; McDonald et al., 1993; Prat-Sala and Branigan, 2000; Christianson and Ferreira, 2005; Altman and Kemper, 2006), definiteness (Grieve and Wales, 1973), and prototypicality (Kelly et al., 1986). What is important here is the fact that, similarly to lexical priming of a referent's name (e.g., Tannenbaum and Williams, 1968; Flores d'Arcais, 1975; Bock and Irwin, 1980; Bock, 1986; Bates and Devescovi, 1989) priming a referent's conceptual accessibility has also been shown to be a strong predictor of Subject selection and the resulting structural choice (e.g., Bock, 1977; Bock et al., 1992; Arnold et al., 2000; Prat-Sala and Branigan, 2000; Christianson and Ferreira, 2005). If conceptual accessibility is related to enhanced memory trace for the corresponding referent's mental representation, then additional memorial activation provided by referent preview should increase the conceptual accessibility of the referent beyond directing attentional focus to it. Hence, the bias to assign the Subject role to the cued referent and to alternate structure accordingly should be particularly strong in cases where a referent preview cue provides information about the cued referent's identity *as well as* points to its location. The effect of a purely visual cue to the location of a referent should, therefore, be weaker because such a cue provides no conceptual information about the referent. An alternative prediction stems from theories that emphasize a special role of attentional focus among non-linguistic factors affecting Subject assignment (Tomlin, 1997; Gleitman et al., 2007; Myachykov et al., 2011). If what matters is only the attentional focus on the cued referent, then there should be no difference in the strength of referential and visual cueing effects. The experiment reported below therefore directly compares the effects of perceptual and referent preview on structural choice. Specifically, we compare the effectiveness of cues that provide only location information with the effectiveness of cues that provide both location *and* referential information.

EXPERIMENT

DESIGN

Two factors were independently manipulated at two levels each: (1) Cue Location (Agent/Patient) and (2) Cue Type (Referent/Dot). Both manipulations were within-subjects and between-items. Cue Location was manipulated by means of presenting a visual cue in the location of one of the subsequently presented visual referents (agent or patient). The dependent variable was the probability of producing Passive-Voice sentences.

PARTICIPANTS

Twenty-four native English speakers (Glasgow University undergraduates; 12 female) with normal or corrected-to-normal vision took part. They either received course credits or £6 subject payment. The mean age of the participants was 20.3 years.

MATERIALS

The target pictures consisted of 64 black-and-white cartoon drawings showing simple transitive events (see example in **Figure 1**) and employed eight different event types (*chase*, *kick*, *pull*, *punch*, *push*, *scold*, *shoot*, and *touch*). Each event type appeared equally often in the Dot-Cue and the Referent-Cue conditions.

The materials were counterbalanced for left–right orientation (i.e., the agent was either on the left or on the right on an equal number of trials), size, animacy, color, and referent role suggestibility (i.e., both referents were equally plausible as being an agent or a patient). The human referents used in the target stimuli appeared in both the agent and the patient role in an equal number of trials. Since it was important that the visual referents were easily recognizable and distinguishable from one another, it was difficult to match them for familiarity. To compensate for this, we provided a practice session at the beginning of each experiment which familiarized participants with all the characters and events they would encounter (see Procedure). The materials were not controlled for corpus frequency; therefore participants previewed the single pictures of each referent during the practice session and became familiarized with the referents they encountered later in the experimental session.

We included 130 filler pictures showing various arrangements of geometrical shapes presented in different regions of the screen (e.g., a square diagonally above and right of a heart); participants had to describe those visual arrangements in the filler trials by producing a locative sentence describing the shapes and the relationship between them. Randomization was constrained so that there were always four fillers at the beginning of each session and each prime–target pair was preceded by at least two filler trials.

APPARATUS

The experiment was implemented in *SR-Research Experiment Builder*. An EyeLink II head-mounted eye-tracker monitored participants' eye movements in order to ensure the efficiency of the cueing manipulation. Other than that, we will not report any eye-movement data since the focus of this paper is on how the experimental manipulations affect speakers' structural choices. The experimental materials were presented on a 17″ CRT monitor of a DELL Optiplex GX 270 desktop computer running at a display refresh rate of 75 Hz. Also connected to the PC was a pair of stereo speakers. A SONY DAT recorder was used for speech recording. The audio clips were later uploaded onto a PC and analyzed with the help of Adobe Audition 2.0. The eye-tracking data were extracted and filtered using *SR-Research Data Viewer*.

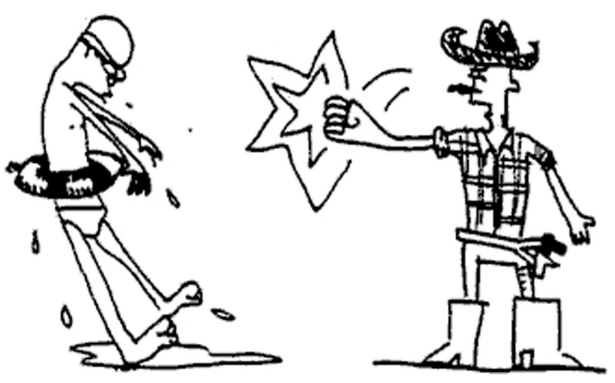


FIGURE 1 | Target-event example.

PROCEDURE

Participants were positioned approximately 60 cm from the display. They had a direct view of the monitor throughout the session. Viewing was binocular, but only the participant's right eye was tracked. Before the main experimental session, each participant was run through a practice session during which they saw the pictures of the referent characters that would later be presented in the target trials and sample pictures of both the target and the filler materials. The referents appeared one at a time in the center of the screen simultaneously with their names. Participants were instructed to read out the referent names and to remember them for the following tasks. Also, each participant had to describe eight sample event pictures (one for each event type) during the practice session. The pictures of the events were presented in the middle of the screen. No specific instruction as to how to describe these event pictures was given to participants, except that participants should always make reference to the event and both interacting characters.

The instruction for the experiment proper was to describe a picture extemporaneously and in a single sentence using the present tense. Participants were unaware of the nature of the experimental manipulations, any difference between target and filler trials, or the exact purpose of the study. They were told that the study was concerned with speaking about what they see on the computer screen. **Figures 2** and **3** illustrate typical target trial sequences.

Each target trial began with the presentation of the central fixation cross. Shortly after the participant fixated it, a dislocated fixation cross appeared on the screen. This ensured that participants would not be looking at the center of the screen at the time of cue presentation and that they would always have to make a saccade to the cued location or, if the cue was overlooked, to another location in the target picture once it appeared on the screen. The dislocated fixation cross was equally distant from the cued locations. The presentation of the cue was contingent on fixating the dislocated fixation mark. Participants fixated the dislocated fixation mark for a minimum of 200 ms, after which either a Dot-Cue or a Referent-Cue screen was displayed. The Dot Cue was a red circle (25 pixels in diameter), which appeared in the approximate center of one of the subsequently presented referents (agent or patient). The Referent Cue was operationalized via previewing one of the event referents (agent or patient) prior to the target display presentation. The previewed referent always appeared in neutral posture, preventing any thematic role (agent or patient) suggestibility. As with the Dot Cues, Referent Cues appeared in the locations corresponding to its location in the subsequently

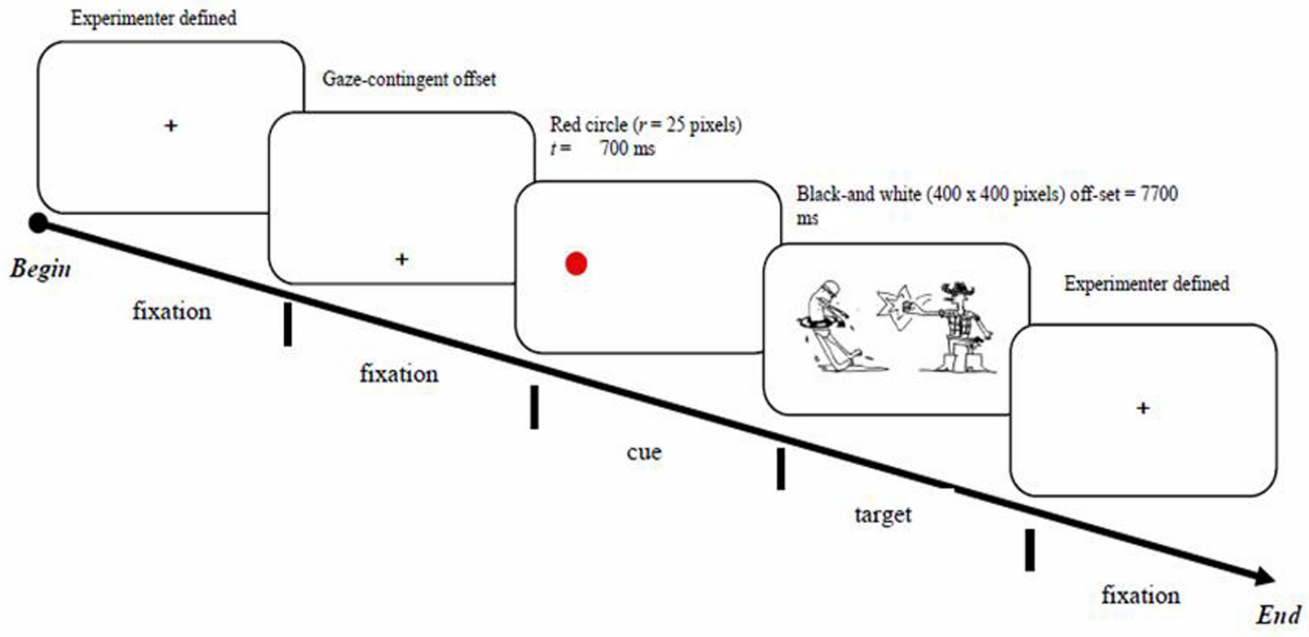


FIGURE 2 | Example of a dot-cue trial.

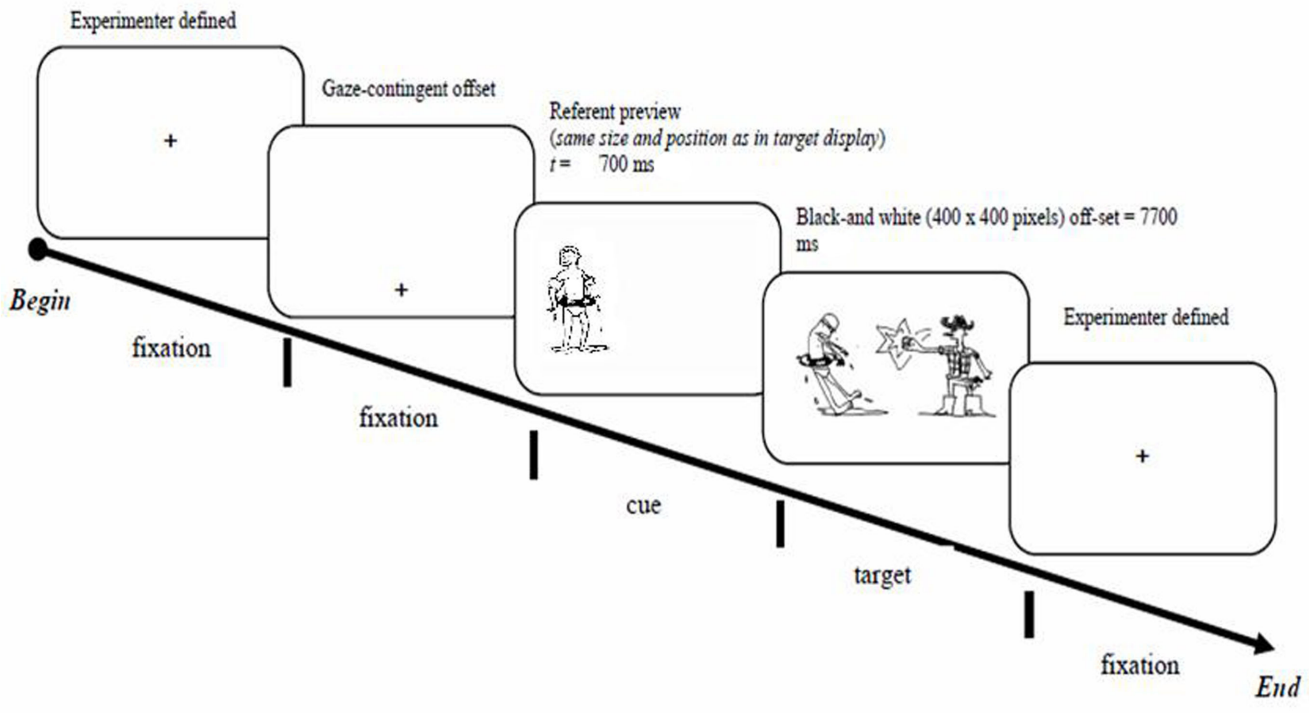


FIGURE 3 | Example of a referent-cue trial.

presented target display. Hence, Dot Cues only provided location information whereas Referent Cues provided both location and referential information. Cue duration was 700 ms regardless of the Cue Type. There was no specific instruction as to how the cues should be treated. After the cue presentation, the target picture appeared on the screen. Participants were instructed to describe

the target picture in a single sentence, and to press the space bar to move on to the next trial. In case the participant did not respond, the picture disappeared from the screen after 7700 ms. Filler trials employed a comparable presentation sequence: the trial would begin with a central fixation mark, after which a dislocated fixation mark appeared, followed by a visual cue (identical to the procedure in the target trials), and finally, the presentation of the target display.

RESULTS

CUEING EFFICIENCY

In order to analyze initial fixations on visually cued versus non-cued referents, the pictures were pre-coded to include separate areas of interest: one for each referent (agent and patient) and one for the background. The referent areas included the referent itself plus a surrounding area of about two degrees of visual angle. Both Dot and Referent cueing manipulations were highly effective in attracting initial visual attention to the cued location. In approximately 96% of the experimental trials, presenting the cue led to the execution of a saccade to the cued location. When the (Dot or Referent) cue was replaced with the target picture (700 ms after cue-onset), participants continued to look at the cued referent, accounting for approximately 90% of initial fixations in the target trials.

TARGET STRUCTURE

Target responses were coded by a naïve coder as Active Voice, Passive Voice, or Other. To be coded as Active Voice, the description had to employ a transitive verb referring to the depicted event, a subject NP referring to the agent, and a direct object NP referring to the patient (e.g., *The cowboy is punching the boxer*). To be coded as Passive Voice, the description had to employ a passivized transitive verb referring to the depicted event, a subject NP referring to the patient, and a by-phrase referring to the agent (e.g., *The boxer is* [*being*] *punched by the cowboy*). Note that truncated passives (not including a by-phrase) were hardly ever produced since they were explicitly discouraged in the practise session. All remaining responses (including missing responses) were coded as Other. The latter accounted for less than 1.5% of the data and will not be considered further.

Statistical analyses were performed in SPSS/PASW 19 using *Generalized Estimating Equations* (GEE, e.g., Hardin and Hilbe, 2003). Unlike ANOVA, GEE allows for specifying distribution and link functions that are appropriate for analyzing categorical frequencies. Here, we used a *binomial* distribution and *log it* link function (cf. Jaeger, 2008) to model proportions of passive-voice responses as a function of Cue Location (agent or patient) and Cue Type (referent or dot). The two predictors were entered as *within-subjects* (respectively *between-items*) variables assuming a compound symmetry covariance structure for repeated measurements. **Table 1** and **Figure 4** present the results of our analysis.

The reliable intercept confirms that passive-voice responses were less likely overall than the active-voice responses. This finding is in line with existing corpus-analysis data (e.g., Svartvik, 1966; Roland et al., 2007) as well as previous findings using visual cueing and referent preview paradigms (see Myachykov et al., 2011 for review). Our analysis registered the presence of a reliable main effect of Cue Location: when the *patient* was cued, passive-voice responses were $23 \pm 10\%$ more likely by subjects and $23 \pm 6\%$ more likely by items, than when the *agent* was cued. This finding provides further support to the previously reported tendency of the attentionally focused referents to correspond to the Subject position in an English transitive sentence. More importantly, in our data there was no suggestion of a Cue-Type effect (dot versus referent) nor of an interaction between Cue Location and Cue Type.

DISCUSSION

In this experiment, we analyzed how directing the speaker's attention to one of the event's referents via prior presentation of a referentially uninformative visual location cue or a referent preview (in the same location) predicts the assignment of the Subject position and the resulting structural choice during English transitive sentence production. Following on from the existing theories, we investigated whether speakers use a combination of perceptual and conceptual information provided by the cue, or only the perceptual information about the location of the cued referent when choosing the Subject and the resulting grammatical structure of a spoken sentence.

In general, cueing the location of the eventual patient resulted in a higher probability of selecting the patient as the sentential Subject and producing passive-voice responses. This finding is in line with previous reports suggesting that attentional focus plays a special role in determining Subject assignment (and corresponding structural choice) in visually situated sentence production (e.g., Tomlin, 1995; Gleitman et al., 2007; Myachykov et al., 2011). The novel finding is that the two different cueing manipulations were equally successful in predicting the choice of Subject, regardless of whether the deployed cue was referentially uninformative or

Table 1 | Results from logit binomial GEE analyses modeling proportions of passive-voice responses as a function of Cue Location (L) and Cue Type (T).

Effect	By subjects		By items	
	GS $\chi^2(1)$	*P*	GS $\chi^2(1)$	*P*
Intercept	37.28	0.001	29.64	0.001
Cue location (L)	13.69	0.001	19.24	0.001
Cue type (T)	0.00	0.949	0.04	0.852
L × T interaction	0.14	0.711	1.13	0.287

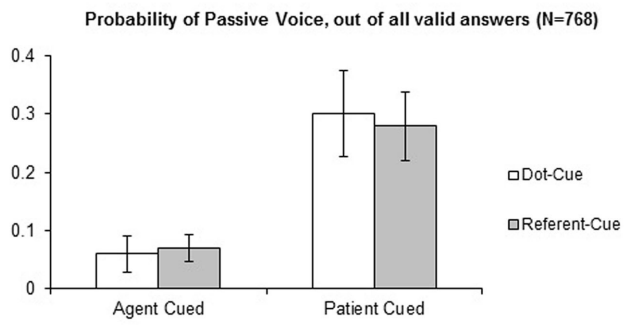


FIGURE 4 | Mean passive-voice probabilities per condition (with by-subject SEs).

was a full referent preview long enough for the speaker to extract both conceptual and lexical information about the cued referent. Contrary to the prediction that conceptual accessibility plays an independent role in determining Subject assignment, the cueing effect on structural choice in the referent-cue condition was no different from that in the dot-cue condition. Our data, therefore, provide further support to the special role of attentional focus in the assignment of the constituents' roles and the resulting structural choice during visually situated sentence production.

Indeed, accessibility-based theories predicted that, in addition to cueing the location of an eventual referent, referent preview would establish a stronger memorial trace for the corresponding referent, which should have led to a further modulation of the overall cueing effect on structural choice. The fact that there was no such modulation might be interpreted as suggesting that participants did not access conceptual and/or lexical information about the previewed referent. This interpretation, however, does not seem very plausible given the amount of time participants were able to preview the referent in the referent-cue condition. It can also be argued that referent preview *alone* is not sufficient to increase the referent's conceptual accessibility, and that other properties of the referent, such as animacy or humanness, need to be manipulated in order to achieve such an increase. This is an interesting empirical question in itself, but its premise takes us back to a very loosely defined notion of conceptual accessibility in the first place. Taking the original definition that "*Conceptual accessibility is the ease with which the mental representation of some potential referent can be activated in or retrieved from memory*" (Bock and Warren, 1985, p. 50), previewing a referent should have achieved exactly that – a better memorial trace for the cued referent's mental representation. If such memorial facilitation played an independent role in Subject selection, then in the design implemented in the current study it should do so *in addition* to biasing attention to one of the subsequently presented referents. The fact that referent preview did not boost the effect of location cueing suggests that attentional focus is the primary driving factor in alternating Subject assignment, thus biasing structural choice.

We propose an alternative interpretation, according to which a stronger memorial representation associated with referent preview plays no additional role in Subject assignment beyond directing attention to the cued referent – the general cueing effect observed in both experimental conditions. In other words, once the speaker commits to using an attentional cue as the predictor of the Subject position, an additional memorial facilitation of the referent-related information does not improve this bias any further. Comparison of our data with the earlier study by Prentice (1967)[2], in which *centrally* established referent preview successfully predicted the assignment of the referent as the Subject, helps to further elaborate our theoretical interpretation. If Prentice's interpretation of her own data was correct in that central referent preview acted as an endogenous cue, orienting participants' attention to the location of the subsequently presented referent, then our study in fact replicates this effect, this time with a lateral referent preview, and using a cue that was mixed: it was both endogenous (that is, prompting participants to identify the previewed referent once the target event was displayed) and exogenous (by virtue of being a laterally presented visual cue accurately predicting the previewed referent's location in the subsequent event; Posner, 1980). The lack of a Cue-Type effect in the current study suggests that referent preview generally acts as a memorial cue to search for the subsequently presented referent. A number of recent reports documented the ability of information held in working memory to affect the distribution of visual attention in perceptual processing tasks (e.g., Bundesen, 1990; Desimone and Duncan, 1995; Downing, 2000; Kumar et al., 2009). In other words, what people currently have in mind can affect what they attend to later. Importantly, these *memorial cues* do not have to be spatial, as linguistic information currently held in working memory has also recently been shown to determine the spatial deployment of visual attention (e.g., Soto and Humphreys, 2007; Hodgson et al., 2009; Mannan et al., 2010; Anderson et al., 2011; Salverda and Altmann, 2011). Our data suggest that once the attentional cue is established in the speaker's working memory, irrespective of whether it was established with the help of a pointer or a referent preview, this attentional cue biases the speaker to select the referent that later appears in the cued location as the sentential Subject. One prediction from this view is that one should observe comparable cueing effects on structural choice for a situation in which, in one condition, referent preview would be established centrally (hence, uninformative about the referent's location), and in the other, laterally (hence, informative about the referent's location). Another way to address this question is via the use of *conflicting* cues, i.e., when a patient referent appears in the agent location (or *vice versa*) at the time of cueing and before the target picture display. This scenario helps address the question of the speaker's selection bias arising from resolving a conflict between information from an endogenous cue (bias to locate the referent whose identity was revealed by the preview) and an exogenous cue (the location of the previewed

[2]It has to be noted that there are important differences between the design used in the current study and the one utilized by Prentice. These include lack of fillers in Prentice (1967), unclear description of the cue slides, and heterogeneous set of referents used in that study: some of the referents were human, others were animals and inanimate objects, and others were indefinite referents, such as *leaves* or *fire*. These features may have affected Subject assignment in their own right.

referent that, in our example, conflicts with its role in the target event).

So, what is the specific role of enhanced memorial activation associated with referent preview in the process of visually situated sentence generation? Our data do not answer this question directly other than suggesting that, as far as Subject selection and structural choice are concerned, there clearly were no cue-enhancing effects of referent preview. However, additional analyses of sentence onset latencies (the time from the onset of the picture to the onset of the participant's response) suggest that participants were on average faster (by 132 ms) to initiate their responses in the referent preview condition than in the dot-cue condition. Although this Cue-Type effect did not approach significance ($ps > 0.1$), the direction of this difference suggests that participants were more prepared to "fill in" the Subject slot by way of pre-activated conceptual and/or lexical access. This interpretation leads to intriguing theoretical implications. It would suggest, for example, that the choice of Subject (which our study showed to depend primarily on attentional focus) is a mechanism separate from the assembly of the corresponding committed structure. Apparently, in both the dot-cue and the referent-cue conditions, participants were biased to assign the Subject role to the referent that later appeared in the cued location. However, in the referent-cued condition, they also knew the identity and the name of the referent, with which they wanted to fill the Subject slot. The difference in sentence onset latency, albeit statistically unreliable, suggests that this knowledge could matter; not at the stage of structural choice, but at the stage of lemma access and linear arrangement of the constituents in the chosen structure. This would be an interesting direction for further research.

In conclusion, we have shown that structural choice (assignment of the Subject role to either the agent or the patient of a transitive event) is primarily driven by attentional factors such as a visual cue to the location of a referent. Additional information about the referent's identity in the cue did not significantly modulate structural choice further, but there might be an influence of referential cueing on conceptual and/or lexical access.

ACKNOWLEDGMENTS

This research was supported by the Economic and Social Research Council grant PTA-026-27-1579 awarded to Andriy Myachykov.

REFERENCES

Altman, L. J. P., and Kemper, S. (2006). Effects of age, animacy and activation order on sentence production. *Lang. Cognitive Proc.* 21, 322–354.

Anderson, S. E., Chiu, E., Huette, S., and Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychol. (Amst.)* 137, 181–189.

Arnold, J., Wasow, T., Losongco, A., and Ginstrom, R. (2000). Heaviness vs. newness: the effects of complexity and information structure on constituent ordering. *Language* 76, 28–55.

Bates, E., and Devescovi, A. (1989). "Competition and sentence production," in *The Crosslinguistic Study of Sentence Processing*, eds B. MacWhinney and E. Bates (New York: Cambridge University Press), 225–256.

Bock, J. K. (1977). The effect of pragmatic presupposition on syntactic structure in question answering. *J. Verbal Learn. Verbal Behav.* 16, 723–734.

Bock, J. K. (1982). Towards a cognitive psychology of syntax: information processing contributions to sentence formulation. *Psychol. Rev.* 89, 1–47.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychol.* 18, 355–387.

Bock, J. K., and Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *J. Verbal Learn. Verbal Behav.* 19, 467–484.

Bock, J. K., Irwin, D. E., and Davidson, D. J. (2004). "Putting first things first," in *The Integration of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. Henderson and F. Ferreira (New York: Psychology Press), 249–278.

Bock, J. K., Loebell, H., and Morey, R. (1992). From conceptual roles to structural relations: bridging the syntactic cleft. *Psychol. Rev.* 99, 150–171.

Bock, J. K., and Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21, 47–67.

Branigan, H. P. (2007). Syntactic priming. *Lang. Linguist. Compass* 1, 1–16.

Bundesen, C. (1990). A theory of visual attention. *Psychol. Rev.* 97, 523–547.

Christianson, K., and Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition* 98, 105–135.

Clark, H. H. (1966). The prediction of recall patterns in simple active sentences. *J. Verb. Learn. Verb. Behav.* 5, 99–106.

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.

Diderichsen, P. (2001). "Selective attention in the development of the passive construction: a study of language acquisition in Danish children," in *Ikonicitet og struktur. Netvork for Funktionel Lingvistik*, eds E. Engberg-Pedersen and P. Harder (Department of English, University of Copenhagen).

Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychol. Sci.* 11, 467–473.

Ferreira, V. S., and Bock, K. (2006). The functions of structural priming. *Lang. Cogn. Process.* 21, 1011–1029.

Flores d'Arcais, G. B. (1975). "Some perceptual determinants of sentence construction," in *Studies in perception. Festschrift for Fabio Metelli*, ed. G. Flores d'Arcais (Milan: Martello-Guinti), 344–373.

Forrest, L. B. (1996). "Discourse goals and attentional processes in sentence production: the dynamic construal of events," in *Conceptual Structure, Discourse and Language*, ed. A. E. Goldberg (Stanford, CA: CSLI Publications), 149–162.

Givon, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics* 30, 5–55.

Gleitman, L., January, D., Nappa, R., and Trueswell, J. (2007). On the give-and-take between event apprehension and utterance formulation. *J. Mem. Lang.* 57, 544–569.

Grieve, R., and Wales, R. (1973). Passives and topicalization. *Br. J. Psychol.* 64, 173–182.

Hardin, J., and Hilbe, J. (2003). *Generalized Estimating Equations.* London: Chapman and Hall/CRC.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends Cogn. Sci. (Regul. Ed.)* 7, 498–504.

Hodgson, T. L., Parris, B. A., Gregory, N. J., and Jarvis, T. (2009). The saccadic Stroop effect: evidence for involuntary programming of eye movements by linguistic cues. *Vision Res.* 49, 569–574.

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506.

Itti, L., and Koch, C., andNiebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* New York, NY: Oxford University Press.

Jaeger, F. T. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446.

Kelly, M., Bock, J. K., and Keil, F. (1986). Prototypicality in a linguistic context: effects on sentence structure. *J. Mem. Lang.* 25, 59–74.

Kumar, S., Soto, D., and Humphreys, G. W. (2009). Electrophysiological evidence for attentional guidance by the contents of working memory. *Eur. J. Neurosci.* 30, 307–317.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation.* Cambridge, MA: MIT.

Mannan, S. K., Kennard, C., Potter, D., Pan, Y., and Soto, D. (2010). Early oculomotor capture by new onsets driven by the contents of working memory. *Vision Res.* 50, 1590–1597.

McDonald, J. L., Bock, J. K., and Kelly, M. H. (1993). Word and world order: semantic, phonological, and metrical determinants of serial position. *Cogn. Psychol.* 25, 188–230.

Myachykov, A., Posner, M. I., and Tomlin, R. S. (2007). A parallel interface for language and cognition in sentence production: theory, method, and experimental evidence. *Linguist. Rev.* 24, 455–472.

Myachykov, A., Thompson, D., Scheepers, C., and Garrod, S. (2011). Visual attention and structural choice in sentence production across languages. *Lang. Linguist. Compass* 5, 95–107.

Myachykov, A., and Tomlin, R. S. (2008). Perceptual priming and structural choice in Russian sentence production. *J. Cogn. Sci.* 9, 31–48.

Osgood, C. E., and Bock, J. K. (1977). "Salience and sentencing: some production principles," in *Sentence Production: Developments in Research and Theory*, ed. S. Rosenberg (Hillsdale, NJ: Erlbaum), 89–140.

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Res.* 42, 107–123.

Pickering, M. J., and Ferreira, V. S. (2008). Structural priming: a critical review. *Psychol. Bull.* 134, 427–459.

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25.

Prat-Sala, M., and Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: a cross-linguistic study in English and Spanish. *J. Mem. Lang.* 42, 168–182.

Prentice, J. L. (1967). Effects of cuing actor vs. cuing object on word order in sentence production. *Psychon. Sci.* 8, 163–164.

Rasolofo, A. (2006). *Malagasy Transitive Clause Types and Their functions.* Ph.D. manuscript, University of Oregon, Eugene.

Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic English grammatical structures: a corpus analysis. *J. Mem. Lang.* 57, 348–379.

Salverda, A. P., and Altmann, G. T. M. (2011). Attentional capture of objects referred to by spoken language. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1122–1133.

Soto, D., and Humphreys, G. W. (2007). Automatic guidance of visual attention from verbal working memory. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 730–737.

Sridhar, S. N. (1988). *Cognition and Sentence Production: A Cross-linguistic Study*. New York: Springer-Verlag.

Svartvik, J. (1966). *On Voice in the English Verb*. The Hague: Mouton.

Tannenbaum, P. H., and Williams, F. (1968). Generation of active and passive sentences as a function of subject or object focus. *J. Verbal Learn. Verbal Behav.* 7, 246–250.

Tomlin, R. S. (1995). "Focal attention, voice, and word order," in *Word Order in Discourse*, eds P. Dowing and M. Noonan (Amsterdam: John Benjamins), 517–552.

Tomlin, R. S. (1997). "Mapping conceptual representations into linguistic representations: the role of attention in grammar," in *Language and Conceptualization*, eds J. Nuyts and E. Pederson (Cambridge: Cambridge University Press), 162–189.

Turner, E. A., and Rommetveit, R. (1968). Focus of attention in recall of active and passive sentences. *J. Verbal Learn. Verbal Behav.* 7, 543–548.

Vigliocco, G., and Hartsuiker, R. J. (2002). The interplay ofmeaning, sound, and syntax in sentence production. *Psychol. Bull.* 128, 442–472.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 September 2011; accepted: 22 December 2011; published online: 18 January 2012.

*Citation: Myachykov A, Thompson D, Garrod S and Scheepers C (2012) Referential and visual cues to structural choice in visually situated sentence production. Front. Psychology **2**:396. doi: 10.3389/fpsyg.2011.00396*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 09 January 2012
doi: 10.3389/fpsyg.2011.00394

Mechanisms and representations of language-mediated visual attention

Falk Huettig[1,2]*, Ramesh Kumar Mishra[3] and Christian N. L. Olivers[4]

[1] *Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands*
[2] *Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, Netherlands*
[3] *Centre of Behavioral and Cognitive Sciences, University of Allahabad, Allahabad, India*
[4] *Cognitive Psychology, VU University Amsterdam, Amsterdam, Netherlands*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Christoph Scheepers, University of Glasgow, UK
Gerry Altmann, University of York, UK

****Correspondence:***
Falk Huettig, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, Netherlands.
e-mail: falk.huettig@mpi.nl

The experimental investigation of language-mediated visual attention is a promising way to study the interaction of the cognitive systems involved in language, vision, attention, and memory. Here we highlight four challenges for a mechanistic account of this oculomotor behavior: the levels of representation at which language-derived and vision-derived representations are integrated; attentional mechanisms; types of memory; and the degree of individual and group differences. Central points in our discussion are (a) the possibility that local microcircuitries involving feedforward and feedback loops instantiate a common representational substrate of linguistic and non-linguistic information and attention; and (b) that an explicit working memory may be central to explaining interactions between language and visual attention. We conclude that a synthesis of further experimental evidence from a variety of fields of inquiry and the testing of distinct, non-student, participant populations will prove to be critical.

Keywords: language, attention, vision, memory, eye movements

INTRODUCTION

A hallmark of human cognition is its ability to integrate rapidly perceptual (e.g., visual or auditory) input with stored linguistic and non-linguistic mental representations. This is particularly apparent during language-mediated eye gaze, a behavior almost all of us are engaged in every day. For instance, when a mother asks her child to "*look at the frog*" or, during dinner, we are asked to "*pass the salt*," linguistic and visual systems, attention and memory processes, must all be quickly integrated. Yet we know surprisingly little about the nature of these cognitive interactions and the representations involved.

How higher level representations involved in language and memory interact with visual input during language-mediated eye gaze has most directly been explored in the *visual world* paradigm in psycholinguistics and the *visual search* paradigm in the field of visual attention. In the visual world paradigm, participants hear an utterance while looking at a visual display (e.g., a semi-realistic scene, or four spatially distinct objects, or printed words; Cooper, 1974; Tanenhaus et al., 1995; see Huettig et al., 2011b, for review). Typically, the display includes objects mentioned in the utterance as well as distractor objects that are not mentioned. The spoken utterances can be instructions to the participants ("direct action" tasks, e.g., "Pick up the candy," Allopenna et al., 1998) or descriptions or comments on the display ("look and listen" tasks, e.g., Huettig and Altmann, 2005). In the latter case, the participants are asked to look at the screen and to listen carefully to the sentences. The participants' eye movements are recorded for later analyses. Some visual world studies have examined whether items that are phonologically, semantically, or visually related (so-called competitors) to a critical spoken word attract attention. Other studies have investigated how the listeners' perception of the scene and/or their world knowledge about scenes and events affect their understanding of the spoken utterances (e.g., whether listeners anticipate up-coming words). In the visual search paradigm, participants are presented with a display of multiple objects and their task is to find a pre-specified target (defined by a certain feature) as quickly as possible (see Wolfe, 1998, for a review). In most studies of these studies, it is assumed that participants will set up some sort of "perceptual" template (or "attentional set") of the target (e.g., when told to "*look for the red square*") for the remainder of the task. The goal of most visual search studies is to investigate the interaction between the bottom-up salience of the stimulus and the top-down goals of the observer (e.g., Treisman and Sato, 1990; Humphreys and Müller, 1993; Wolfe, 1994; Cave, 1999; Itti and Koch, 2000; Palmer et al., 2000). An important difference between the two paradigms is that in the visual world paradigm the visual display precedes (or occurs simultaneously) with the spoken instruction (or sentence) whereas in visual search studies the (linguistic or visual) instruction precedes the search display.

In short, the main interest of researchers using the visual world paradigm tends to be on aspects of linguistic processing whereas visual search investigators are primarily interested in what determines the efficiency of the search process, how easily conjunctions of basic features (e.g., color and shape) can be found, and whether search involves serial or parallel processing. These distinct focal points of interest have resulted in a theoretical no-man's land in which the exact nature of the interaction of linguistic and visual processing, and of attention and memory, have been left largely unexplored.

The aim of the present paper is to highlight, (a) theoretical challenges to explaining how language, vision, memory, and attention interact and, (b) empirical challenges in view of recent data with young children and illiterates/low literates in the visual world paradigm. We argue that existing theoretical proposals do not discuss (at all or in sufficient detail) four major underpinnings of this oculomotor behavior: levels of representation involved, attentional mechanisms, the nature of memory, and the degree of individual and group differences.

We will therefore first discuss the levels of representation at which language-derived and vision-derived representations are integrated (see Levels of Representation). An explanation of attention will be central for a mechanistic account about how this oculomotor behavior is instantiated and thus, in Section "Attention," we consider the attentional mechanisms which may underlie language-mediated eye gaze. Language–vision interactions of course also involve temporary and long-term memory storage; we reflect on what types of memory may be involved and their nature (see Memory). In Section "Individual and Group Differences," before concluding, we discuss empirical challenges for the investigation of the mechanisms and representations shared by language, vision, attention, and memory; in particular the need to study distinct, non-student, participant populations.

LEVELS OF REPRESENTATION

To understand how language interacts with vision, it is necessary to establish what knowledge types are retrieved when someone is confronted with both language and visual input, as well as how these linguistic and visual representations interact. Furthermore, such representations are likely to change over time as the linguistic input unfolds and the visual image has been available for some time. An early linguistic–visual linking hypothesis was proposed by Tanenhaus and collaborators (Allopenna et al., 1998) which Huettig and McQueen (2007) termed the *phonological mapping hypothesis*. According to this hypothesis, phonological representations are activated by both spoken words and visual objects (i.e., the names of the objects in the display). A match in phonological representations retrieved from both modalities results in an increased likelihood of a saccade toward the location of the (partially) matching visual source. This is in line with many models of spoken word recognition which assume that at a phonological level different candidate words are considered in parallel (cf. Marslen-Wilson and Welsh, 1978; Marslen-Wilson, 1987). Continuous mapping models of spoken word recognition (e.g., McClelland and Elman, 1986; Gaskell and Marslen-Wilson, 1997) assume that lexical access during spoken word recognition is continuous and thus predict that rhyming words (e.g., beaker/speaker) should also be at least weakly activated.

Consistent with these models, Allopenna et al. (1998) observed that the likelihood of fixations to both a picture of a beaker and a picture of a beetle increased as participants heard the word "*beaker*." As acoustic information from "*beaker*" started to mismatch with the phonological information of "beetle," the likelihood of looks to the beetle decreased as the likelihood of looks to the beaker continued to rise. In addition, looks to a picture of a speaker started to increase as the end of the word "*beaker*" acoustically unfolded. The finding that simulations with the TRACE model (McClelland and Elman, 1986) of speech perception, which includes an explicit phoneme layer, closely fit the eye movement data of Allopenna et al. (1998) provided further support for the phonological mapping hypothesis.

It is however important to note that the many demonstrations of the influence of acoustic–phonetic information in visual world studies (e.g., McMurray et al., 2002; Salverda et al., 2003; Shatzman and McQueen, 2006) are consistent with the phonological mapping hypothesis but do not necessarily provide support for it. This is because there is general agreement that spoken word recognition is a cascaded rather than a strictly serial process (e.g., that information from the acoustic signal cascades to higher levels before processing at lower levels is completed) and that thus activation of word form representations cascades further to, for instance, morphological, semantic, and syntactic representational levels. Thus, the initial phonological representations retrieved on hearing the spoken word "*beaker*" may activate semantic representations of beakers as well as beetles, and the mapping between spoken words and the different competing visual objects may therefore take place at the level of semantic/conceptual rather than phonological (or phonetic) representations. This is the *semantic mapping hypothesis*. One could go even further than that and argue that activation of phonetic and semantic representations automatically spreads to the associated visual shapes and thus the match with the visual input occurs at a perceptual level. We could call this the *visual mapping hypothesis*.

Semantic mapping effects were first reported by Cooper (1974), who observed that participants were more likely to fixate pictures showing a snake, a zebra, or a lion when hearing the semantically related word "Africa" than they were to fixate referents of semantically unrelated control words. However, Cooper did not investigate systematically the nature of the semantic effects he observed (e.g., the words "Africa" and "lion" are not only semantically but also associatively related, as they often co-occur, like "computer" and "mouse"). Huettig and Altmann, 2005, see also Yee and Sedivy, 2001, 2006; Dunabeitia et al., 2009) further pursued Cooper's finding by investigating whether semantic properties of spoken words could direct eye gaze toward objects in the visual field in the absence of any associative relationships. Huettig and Altmann (2005) found that participants directed overt attention toward a depicted object (e.g., a trumpet) when a semantically related but not associatively related target word (e.g., "*piano*") acoustically unfolded, and that the likelihood of fixation was proportional to the degree of conceptual overlap (cf. Cree and McRae, 2003). In a similar study (Huettig et al., 2006; see also Yee et al., 2009) observed that several corpus-based measures of word semantics (latent semantic analysis, Landauer and Dumais, 1997; contextual similarity, McDonald, 2000) each correlated well with fixation behavior. Thus, language-mediated eye movements are a sensitive indicator of the degree of overlap between the semantic information conveyed by speech and the conceptual knowledge retrieved from visual objects. The fact that phonological relationships were avoided between spoken words and visual objects in the semantic studies shows that semantic mapping behavior can occur in the absence of phonological mapping.

Evidence for visual mapping (i.e., increased looks to visually related entities, e.g., matching in color or shape) have also been

observed. For example, participants shifted overt attention to a picture of a cable during the acoustic unfolding of the word "*snake*" (shape being the obvious match here, Huettig and Altmann, 2004, 2007; Dahan and Tanenhaus, 2005). In a related study, Huettig and Altmann (2004) found that participants shifted their eye gaze to a picture of a strawberry when they heard "*lips*" (presumably on the basis of the typical color of these objects). The likelihood of fixating a particular visual object thus reflects the overlap between stored knowledge of visual features of a word's referent, accessed on hearing the spoken word, and visual features extracted from the objects in the visual environment.

It is important to note that there are two possible ways in which visual mapping may occur: between the typical visual form of the referent retrieved on hearing the spoken word (e.g., the typical shape of snakes on hearing "*snake*" or the typical color of lips on hearing "*lips*") and the *perceived* visual form or color of the displayed object (in absence of any stored visual form object knowledge) and/or the stored *knowledge* about the typical visual form or color of the displayed object (as retrieved from viewing the object). The shape of an object, the long and thin form of a cable, or the color of a strawberry, can be perceived but is also known. Eye movements that are contingent upon currently perceived information (which may be temporarily stored in visual working memory) cannot easily be dissociated from eye movements that are contingent upon stored information about object form (see also Yee et al., 2011). To investigate this issue Huettig and Altmann (2011) manipulated the presence of color in a series of experiments. The *conceptual* representation of an object's color (i.e., the stored color knowledge about an object) and the *perceived* but non-diagnostic color of an object (i.e., its surface color) can be dissociated. Participants were presented with spoken target words whose concepts are associated with a typical color (e.g., "*spinach*") while their eye gaze was monitored to (i) objects associated with the same typical color but presented in black and white (e.g., a black and white line drawing of a frog), (ii) objects associated with the same typical color but presented in an appropriate but atypical color (e.g., a color photograph of a yellow frog), and (iii) objects typically not associated with the color but presented in the color associated with the target concept (e.g., a green blouse). No effect of stored object color knowledge was found when black and white line drawings or black and white photos were used. A small effect of stored object color knowledge was found when color photographs were used depicting the target object (e.g., a frog) in an atypical but appropriate color (e.g., a yellow frog). The finding that the effect was marginal and occurred more than 1 s after information from the acoustic target word started to become available suggests that stored object color, if anything, has a minor influence on language-mediated eye movements. In contrast, Huettig and Altmann (2011) found a large bias toward objects displayed in the same surface color (as the prototypical color associated with the spoken word) even though the referent of the picture (e.g., a green blouse) was not itself associated with that color. These experiments suggest that online visual mapping between spoken words and visual objects is mainly contingent upon the perceived visual information (temporarily stored in visual working memory) rather than stored object form or color knowledge accessed on viewing the visual objects. Overall thus, three main hypotheses (visual, phonological, and semantic mapping) about the representational levels at which linguistic and visual input match have been proposed. Some recent research has been directed at evaluating these hypotheses.

To counter criticism that looks to phonological competitors in the visual world paradigm might just be due to strategic covert object naming rather than normal lexical analysis of the spoken words (i.e., that the phonological effects reflect a match between the phonological input of the spoken words with strategically retrieved object names bypassing further lexical analysis of the spoken words), Dahan and Tanenhaus (2005) have recently argued that the visual competition effects are "inconsistent with the hypothesis that eye movements merely reflect a match between the unfolding speech and pre-activated phonological representations associated with object locations" (p. 457). They then go on to claim that mapping occurs at the perceptual level, not the lexical level. This is correct in the sense that visual (and semantic) effects also occur in absence of phonological overlap, ruling out the claim that "word–object matching" in the visual world paradigm is *entirely* due to phonological mapping. The visual effects however do not rule out that phonological and semantic mapping (at least sometimes) occur. Moreover, from word–object mapping at a phonological level of representation does not necessarily follow that there is no further lexical analysis of the spoken words.

There is evidence from other paradigms showing that viewers often access the names of objects, even when they do not intend to name them (e.g., Noizet and Pynte, 1976; Zelinsky and Murphy, 2000; Morsella and Miozzo, 2002; Navarette and Costa, 2005; Meyer and Damian, 2007; Meyer et al., 2007; Mani and Plunkett, 2010). Noizet and Pynte (1976) for instance asked their participants to shift eye gaze to three objects, one after another, and to identify them silently. Participants were told that no response was required and that they would not be tested afterward. Noizet and Pynte (1976) observed that participants gazed about 200 ms longer at objects with multi-syllable names (e.g., hélicoptère) than objects with one-syllable names (e.g., main; see Zelinsky and Murphy, 2000, for a similar result). Morsella and Miozzo (2002) used a picture–picture version of the Stroop task in which speakers were shown pairs of superimposed pictures and were instructed to name one picture and ignore the other. They found that participants were faster at naming pictures with distractors that were phonologically related. Thus, the pictures participants were instructed to ignore exerted a phonological influence on production which suggests that participants retrieved the phonological forms of the names of the distractor pictures. As a final example, Mani and Plunkett (2010) recently showed that even 18-months-olds implicitly name visual objects and that these implicitly generated phonological representations prime the infants' subsequent responses in a paired visual object spoken word recognition task. These results suggest that viewing a display of visual objects does result in lexical analysis of the displayed objects, at least in these tasks and if participants have sufficient time to inspect the scene/display.

Huettig and McQueen (2007) tested the hypothesis that neither the simple phonological or visual or semantic mapping hypotheses are correct and that instead there appears to be a complex three-way tug of war among matches on all three levels of representation. In four experiments, participants listened to spoken

sentences including a critical word. The visual displays contained four spatially distinct visual items (a phonological, a semantic, and a visual competitor of the critical spoken word, and a completely unrelated distractor). When participants were given sufficient time to look at the display (i.e., before the critical spoken word), shifts in eye gaze to the phonological competitor of the critical word preceded shifts in eye gaze to shape and semantic competitors. Importantly, with only 200 ms of preview of the same picture displays prior to onset of the critical word, participants no longer preferred the phonological competitor over unrelated distractors, and prioritized the shape and semantic competitors instead. Thus it appears that when there is plenty of time to view the display picture processing progresses as far as retrieval of the pictures' names. But when there was only 200 ms of preview before the onset of the critical spoken word, picture processing still involved retrieval of visual and semantic features, but there was insufficient time to retrieve the pictures' names.

Yee et al. (2011) have recently suggested that *long-term* knowledge about an object's form becomes available *before* information about its function (cf. Schreuder et al., 1984; but see Moss et al., 1997) based on their finding that eye movements mediated by conceptual shape (i.e., a slice of pizza activating the round shape of a whole pizza) were observed with 1000 ms but not with 2000 ms preview of the visual display. The opposite pattern was observed for looks to semantic competitors (i.e., no effects with 1000 ms but a significant bias with 2000 ms preview). This pattern of results is striking but the semantic results appear to be inconsistent with previous research since strong semantic effects have been observed with as little as 200 ms preview in other visual world studies (see Table 4 of Huettig and McQueen, 2007; see also Dell'Acqua and Grainger, 1999, for evidence that 17 ms exposure to pictures of objects is enough to activate gross semantic category information). Future studies could usefully be directed at investigating the differences underlying these seemingly contradictory results.

There is evidence that the nature of the visual environment induces implicit biases toward particular types of mapping during language-mediated visual search. This is because Huettig and McQueen (2007) found a different pattern of results when the pictures were replaced with printed words (the names of the same objects as before). Under these conditions shifts in eye gaze were directed only to the phonological competitors, both when there was only 200 ms of preview and when the displays appeared at sentence onset. This suggests that eye gaze is co-determined by the type of information in the display (i.e., visual objects or words). Further support for this notion was provided in a subsequent series of experiments (Huettig and McQueen, 2011). The same sentences and printed words as in Huettig and McQueen (2007) were used. When semantic and shape competitors of the targets were displayed along with two unrelated words, significant shifts in eye gaze toward semantic but not shape competitors were observed as targets were heard. The results were the same when, semantic competitors were replaced with unrelated words, and in addition, semantically richer sentences were presented to encourage visual imagery, and moreover, participants rated the shape similarity of the stimuli before doing the eye-tracking experiment. Yet none of the cases resulted in rapid shifts in eye gaze to shape competitors, There was a late shape-competitor bias (more than 2500 ms after target onset) in all experiments, which shows that participants can in principle access shape information from printed words. These data thus show that shape information is not used in online search of printed word displays whereas it is used with picture displays. In other words, the likelihood of mapping between language-derived visual representations and vision-derived visual representations is contingent upon the nature of the visual environment. Finally, at least when printed word displays are used, recent results suggest that language–vision mapping can also occur at an orthographic representational level (Salverda and Tanenhaus, 2010; see also Myachykov et al., 2011, for discussion of mapping processes at the syntactic level in a language production task; and Mishra and Marmolejo-Ramos, 2010, for an embodied cognition account).

In sum, research has shown that with picture displays, fixations can be determined by matches between knowledge retrieved on the basis of information in the linguistic and in the visual input at phonological, semantic, and visual levels of representation. With printed word displays, fixations are determined by online matches at phonological, semantic, and orthographic levels. The exact dynamics of the representational level at which such mapping occurs however is co-determined by the timing of cascaded processing in the spoken word and object/visual word recognition systems, by the temporal unfolding of the spoken language, and by the nature of the visual environment (e.g., which other representational matches are possible).

ATTENTION

The mapping hypotheses outlined so far describe the levels at which language-derived and vision-derived representations match during language-mediated eye gaze. They do not however provide any mechanistic account about how this oculomotor behavior is instantiated. Attention will probably be central to such an explanation, as the eye movements are likely an overt expression of shifts in the attentional landscape (such shifts may of course also occur covertly, e.g., Posner, 1980). Within the field of attention research, objects in the visual field are assumed to compete for representation, with the strongest object being selected for further behavior (e.g., a manual or oculomotor response; Wolfe, 1994; Desimone and Duncan, 1995; Itti and Koch, 2000; Miller and Cohen, 2001). This competition is generally thought to be biased by two types of mechanism: a bottom-up or feedforward mechanism representing stimulus strength, and a top-down or feedback mechanism representing the current goals of the observer (see, e.g., Theeuwes, 2010, for a review). For example, a bright red poppy in a field of grass may automatically capture one's eyes, but it will especially do so if one is looking to compile a nice bouquet of wild flowers.

Note that this attentional framework is not immediately applicable to visual world behavior. For one, in many visual world studies there is no clear task goal that would *a priori* be expected to induce visual biases. The task is often simply to look around and at the same time to just listen to the spoken input. As has been pointed out recently (Huettig et al., 2011b; Salverda et al., 2011) visual world type interactions may well be modulated by different task settings, but so far this has received little systematic investigation. Furthermore, visual world experiments are typically little concerned with the visual stimulus properties. The visual objects are chosen for linguistically relevant characteristics (i.e., their names

or meanings), and not their physical characteristics (though see Huettig and Altmann, 2004, 2007, 2011; Dahan and Tanenhaus, 2005).

We can remedy this by assuming that not only visual features or task goals add to the attentional weight of a visual object, but also its linguistic (e.g., phonological) and semantic properties. Indeed this is what a number of models reported in the visual world literature do. Roy and Mukherjee's (2005) probabilistic rule model integrates sentence-level and visual information, such that each word in an unfolding sentence incrementally influences the distribution of probabilities across the visual scene, based on the fit of the visual context with the current word. The distribution of probabilities are interpreted as attentional distributions, such that processing priority is assumed to be distributed over the visual objects in the scene. According to Altmann and colleagues (Altmann and Kamide, 2007; Altmann and Mirkovic, 2009), attending to a language-matching visual object is an emergent property of spreading activation. The visual and linguistic input overlap at for example the semantic level, where they reinforce each other. This increased activation then spreads back to the specific linguistic and visual representations, including the visual location, which then serves as a saccadic target. In the model of Mayberry et al. (2009), attention is directed to identified visual regions in order to establish a reference for the spoken input. The relationship between language and vision is reciprocal, in that the referent (i.e., attended) object in turn influences the interpretation of the incoming speech. In other words, the language comprehension system makes use of whatever information is available, including visual information. This way, language becomes grounded in a visual environment, in line with for example developmental findings. Likewise, a neural net implementation of the model learns to interpret ambiguous linguistic input by attending to seemingly relevant (i.e., matching) visual input. The net result is the same as for the other models: matching visual input becomes more strongly represented. Finally, in Kukona and Tabor's (2011) recent dynamical systems model of the visual world paradigm, attention is expressed as a landscape of local attractors reflecting the visual objects, a landscape that continuously changes on the basis of the linguistic input.

Whereas psycholinguistics has welcomed attention into their models, very few visual search studies have looked at the role of language. One exception is a study by Wolfe et al., 2004; see also Vickery et al., 2005), who compared visual search under verbal (i.e., written) instructions to that under visual instructions. Observers were asked to search a complex display for a unique (but non-salient) target. The target changed from trial to trial, as was indicated by an instruction. This instruction was either pictorial in nature (i.e., it showed an exact picture of the target), or it was a written description (e.g., it read "blue square"). Furthermore, the SOA between the cue and the search display was varied. The results showed that pictorial cues were very effective: already for SOAs of 200 ms, performance reached asymptote, and search was as fast as in a baseline condition in which the target always remained identical from trial to trial (and thus no instruction was necessary). Performance was considerably worse for the written cues. Search speed was never comparable to the baseline condition, and even after 1600 ms (the greatest SOA measured) it had not reached asymptote yet. This despite the fact that the written cues described very simple visual forms that the observers had seen over and over again during the course of the experiment. This suggests that, in visual search, observers do not necessarily create a visual template from a verbal description, and instead complete the task on the basis of a less precise representation which could be linguistic in nature, but is in any case more abstract than a visual template.

Whatever the precise model, note that for linguistic content to be translated into a spatial attentional landscape, a considerable binding problem needs to be solved, linking the phonological and semantic codes to a specific visual location. Cognition needs what has been referred to as *grounding*, *situating*, or *indexing*. This problem has been recognized by many (e.g., Richardson and Spivey, 2000; Kukona and Tabor, 2011), but so far has not been adequately solved by visual world models. According to Altmann and Mirkovic (2009), the increased activation of the overlapping representations within a supramodal network automatically spreads back to the matching object's location. Such a network is not necessarily a separate supramodal module in itself, but may emerge from the global, linked activity in the range of networks involved in representing the visual and linguistic input. Useful as it is as a general explanatory framework, it begs the question as to how a representation within such a network knows what the (spatial) source is of its activity. If everything is active, how can one piece of information be specifically bound to another? In the typical visual world display, there are multiple objects, and hence multiple active locations, any of which could be the source. Altmann and Mirkovic propose that an object's location as well as its more symbolic properties are part of one and the same "representational substrate," but they left unspecified how this representational substrate would look like.

The problem has been recognized within the attention literature, where the question boils down to how separate visual features such as color and orientation can be tied to a specific object or location (Treisman and Gelade, 1980; Treisman, 1996; Reynolds and Desimone, 1999). One classic solution has been the idea that by locally attending to an object, its features will become activated together. Thus, attention causes binding. Obviously, this solution does not suffice here, since we try to explain exactly the opposite: how the binding of information causes attention. One promising way of creating a representational conglomerate that includes an object's location as well as its identity is through local interactions of feedforward and feedback mechanisms (e.g., Lamme and Roelfsema, 2000; van der Velde and de Kamps, 2001; Hamker, 2004; Vanduffel et al., 2008). The idea is that a visual target object is first represented in low-level perceptual layers, which due to their retinotopic organization and small receptive fields include detailed spatial information. These layers then feed forward into layers that eventually recognize the identity of the object. These higher layers are not retinotopically organized and due to large receptive fields, location information is largely lost. Part of the recognition layers will recognize the target object and become active accordingly. This activity is fed back to the lower layers, but due to the loss of location information this feedback is spatially non-specific. However, the feedback can be made spatially specific by making it interact with the feedforward activity that drove the recognition in the first place. That is, at each layer, the feedback is gated by,

or correlated with the feedforward activity that fed into that layer. Thus, the feedback trickles down the representational ladder and becomes more and more localized, thus tying a recognition unit to a specific visual instantiation. There is no *a priori* reason why layers representing linguistic information about visual objects could not be linked in the same fashion, and thus create the representational substrate proposed by Altmann and Mirkovic (2009).

In sum, little research so far has investigated the exact nature of the attentional mechanisms underlying language-mediated eye gaze. The most concrete proposal to date postulates that language-mediated visual orienting arises because linguistic and non-linguistic information and attention are instantiated in the same common coding substrate. Local microcircuitries involving feedforward and feedback loops may instantiate such a representational substrate.

MEMORY

As with virtually any cognitive process, the interactions between language and eye movements involve memory. The question is what types of memory are involved. There is no doubt that long-term memory plays a crucial role, as it provides the semantic, phonological, and visual knowledge base (or "type" representations) on which these interactions are based. Spreading activation then travels along the associations formed within and between these different types of knowledge networks. Indeed there is growing evidence that both visual and semantic knowledge stored in long-term memory representations automatically affect visual selection. For example, in a visual search task, Olivers (2011) asked participants to search a display for a grayscale version of a known traffic sign. On each trial a distractor sign was presented in a color which was either related or unrelated to the target sign. For example, when looking for a black and white hexagonal STOP sign (which is usually red in Europe) the distractor could be a red triangular warning sign (related) or a blue square parking sign (unrelated). Distractors interfered more with participants' search when the color of the distractor sign was related than when their color was unrelated even though color was completely irrelevant to the task. Apparently, the participants could not help but retrieve the associated color. Similarly, Moores et al. (2003) found interference stemming from a conceptual relationship. For example, when observers were asked to look for a picture of a motorbike, they were more distracted by a picture of a helmet than a picture of a football. Finally, Meyer et al. (2007) reported interference from an overlap in object name, for example when observers were asked to look for a bat (the animal), they were distracted by a picture of a baseball bat. Similarly, Soto and Humphreys (2007) found that after the instruction to remember the word "red," observers were more distracted by red objects in the display. Although some working memory was involved in this study, the link between the word and the visual color representation must obviously rely on LTM knowledge.

However, as argued earlier, the mere spread of activation on the basis of long-term links is insufficient to explain such findings in visual search, as well as visual world behavior. Note that both visual search and visual world displays are often characterized by a substantial degree of arbitrariness in the collection of objects presented and the locations where these objects are put. Unlike real world scenes in which particular objects are often associated with particular locations (for example when opening the fridge, the milk bottle is typically located in the lower door compartment), in visual world displays the target object (e.g., the "trumpet") may be presented in the top left of the screen on one trial, and in the bottom-right on the next. There is no *a priori* long-term memory that links these objects to those locations, yet attention is directed there. Some temporary memory therefore seems necessary, a memory that links the type representations to a "token" representation of the specific instance of an object in a spatiotemporal world (also referred to as object files, indices, or deictic pointers; Kanwisher, 1987; Kahneman et al., 1992; Pylyshyn, 2001; Spivey et al., 2004; Hoover and Richardson, 2008).

The nature of this temporary memory is subject to debate. Some refer to it as being "episodic" (e.g., Altmann, 2004; Altmann and Kamide, 2007), but that obviously says little about its exact nature. The field will need to answer questions such as whether the binding of linguistic types to visual tokens is an implicit process, occurring automatically, without much cognitive control and/or awareness, or an explicit process, relying on the awareness of the stimuli involved, and therefore subject to cognitive control but also capacity limitations. Implicit representations are more likely to last for a longer period, while shorter term explicit memories are more subject to interference. Naturally, both types of memory may contribute to visual–linguistic interactions. An implicit memory is most clearly advocated by Altmann and colleagues (Altmann and Kamide, 2007; Altmann and Mirkovic, 2009), who argue that visual world type interactions are inevitable given the automatic spread of activation within a conglomerate of linguistic and visual representations. As we have argued above, such an account could work if the sprawl of activity can be channeled back to the original source – something that can be achieved through gating the feedback signal with the feedforward signal between layers of representation (van der Velde and de Kamps, 2001). Another argument for an implicit mechanism is that visual world interactions occur even though the visual and spoken input are often irrelevant to the observer (i.e., there is no explicit physical task), suggesting a substantial automatic component.

Others have advocated an important role for an explicit type of memory, most notably working memory (Spivey et al., 2004; Knoeferle and Crocker, 2007; Huettig et al., 2011a). The fact that visual world effects occur despite the absence of a clear task does not preclude such a contribution. After all, participants are at the very least instructed to "just" look at the display and "just" listen to the input, which may facilitate at least a partial entrance into working memory. One reason for assuming this type of memory comes from visual attention studies that suggest that the number of visual tokens or indices that can be simultaneously maintained is limited to four – a limit assumed to be the limit of visual working memory (Cowan, 2001). If visual world interactions depend on such tokens, they would thus also depend on visual working memory. But also on the psycholinguistic side, it has been argued that working memory is a real prerequisite for disambiguating and understanding language (Jackendoff, 2002, see also Marcus, 1998, 2001). It remains to be tested whether visual world effects are also subject to a limit of four visual objects and how they respond to different forms of cognitive load.

One advantage of the explicit memory account is that it allows the cognitive system to flexibly juggle the maintenance of visual memories between the internal and external world. As long as a visual stimulus is present, in principle it suffices to have only a minimal visual memory representation of them. Instead, the indices or pointers can be used to refer to the location of the object, allowing the cognitive system to only retrieve detailed percepts when necessary. This way the world serves as an outside memory, limiting the load on the cognitive system (O'Regan, 1992; O'Regan and Noë, 2001; Spivey et al., 2004). This would mean that the spatial pointers as alluded to when explaining visual world type effects are not just side effects of a memory system that cannot help but bind all sorts of information, but actually have a functional role in establishing the memory in the first place by directly referring to the outside world (a reference that then may be sustained even if the outside scene has been removed). A study by Wolfe et al. (2000) is directly relevant here. In some of their experiments, they presented observers with a visual search display that remained constantly on screen from trial to trial. The specific target changed from trial to trial (through an instruction in the center of the screen). For example, the search display might always consist of a red circle, a green square, a red triangle, and a blue diamond – all continuously present in the same position. On the first trial the target may then be a green square, whereas on the next it may be the red circle, and so on. Remarkably, even though the search display remained constant from trial to trial, search hardly improved. Even after 300 trials there was no notable improvement in search. Wolfe et al. (2000) concluded that no memory of the display was built up, despite countless inspections. They argued that for the lazy cognitive system, learning the display was unnecessary, since the stimulus remained visible and could be used as an outside memory. In contrast, when the search display was taken away after the first presentation, performance rapidly became fast and efficient. Now observers were forced to commit the items to internal memory, making them more rapidly available for selection. This flexibility (as induced by task demands) suggests some form of working memory, but it remains to be seen whether visual world interactions are equally flexible.

That working memory content *can* guide visual attention has been shown in several studies now (Soto et al., 2005; Olivers et al., 2006; Soto and Humphreys, 2007; Olivers, 2009). In these studies, observers are asked to look for a simple visual shape target among distracters, while keeping an unrelated object in working memory. However, one of the search distracters can match the memorized object (e.g., in color), and when it does, search suffers. It appears that an object that matches the contents of working memory captures attention, something which has been confirmed with eye movement measures. Of course, the fact that working memory can affect attentional guidance does not necessarily mean that it also does so in visual world settings. This remains to be investigated (see Huettig et al., 2011a, for a more detailed review).

INDIVIDUAL AND GROUP DIFFERENCES

The vast majority of studies investigating language-mediated eye gaze have been conducted with undergraduate students. This is of course not only the case for studies using the visual world and visuals search paradigms but a pervasive problem in experimental psychology more generally (see Arnett, 2008). It is an open empirical question how much one can generalize from the sophisticated behavior of highly educated university students to draw general inferences about mind and behavior beyond these narrow samples. Indeed it has been argued that the homogeneous Western student participants used in most studies are the "weirdest" (Western Educated Industrialized Rich Democratic) people in the world and the least representative populations one can find to draw general conclusions about human behavior (Henrich et al., 2010, for further discussion). Besides the theoretical challenges discussed above, there are thus some empirical challenges, which research on the interaction of the cognitive systems involved in language, vision, attention, and memory, must address. One promising line of inquiry will be the investigation of individual differences (see McMurray et al., 2010, for an example). Another approach, and one we shall discuss here in more detail, are studies with distinct non-student participant populations. Recent studies investigating language-mediated visual orienting in young children and in individuals with little formal schooling (i.e., low literacy levels) suggest that this approach may prove to be particularly fruitful.

There is the possibility that the mapping between spoken words and visual objects is mediated by stored verbal labels. Consider the color effects reported by Huettig and Altmann (2004, 2011). On hearing target words that are associated with a prototypical color (e.g., "*frog*"), participants tend to look at objects displayed in that color even though the depicted objects (e.g., a green blouse) are not themselves associated with that prototypical color (see Johnson and Huettig, 2011, for a similar results with 36-month-olds). But when listeners hear the word "*frog*," do they access an associated stored color label (GREEN), which makes them more likely to look at green things in their visual surroundings? Or, alternatively, do listeners on hearing "*frog*" access a target template, a sort of veridical perceptual description of the target (including its color) which then leads to a match with items matching this "perceptual" template (as tends to be assumed in visual search studies)? Note that verbal mediation is a genuine possibility; participants in free word association tasks typically produce the answer "green" when asked to write down the first word that comes to mind when thinking about "frog" (Nelson et al., 1998). Davidoff and Mitchell (1993) for instance have argued that "3-year-olds have more difficulty matching object colors with mental templates than they do with color naming" (p. 133) based on the finding that their 3-year-old participants tended to successfully judge that a banana is colored yellow in a verbal task but failed to choose the yellow banana as the correct one from differently colored bananas. Moreover, developmental psychologists have argued that "early in life, sensory, and linguistic color knowledge seem to coexist, but a proper map connecting names and perception is late in developing" (p. 78, Bornstein, 1985).

To examine this issue, Johnson et al. (2011) tested 48 two-year-olds who lacked reliable color term knowledge and found that on hearing the spoken target words they looked significantly more at the objects that were either color-related or semantically related to the named absent targets (e.g., on hearing "frog" they were more likely to look at a green truck and a bird than completely unrelated objects). Interestingly, there was a clear dissociation: words such as "*frog*" resulted in shifts in eye gaze to green things but color

words such as "*green*" did not. Thus, 2-year-olds look to color-matched competitors even if they do not know the label for that color. The Johnson et al. (2011) results do not rule out that adults have both direct and indirect routes linking color knowledge of words. What the Johnson et al. (2011) results suggest, however, is that the direct perceptual route exists before the indirect, lexically mediated route, has had a chance to develop.

Recent research involving adult individuals with little formal schooling also provides new insights with regard to the mechanisms and representations during language-mediated visual orienting. Studies using the blank screen paradigm (Spivey and Geng, 2001; Altmann, 2004), in which participants preview a visual scene and then listen to a spoken sentence while a *blank* screen is shown, have found that people have a tendency to re-fixate the regions on the blank screen that were previously occupied by relevant objects. Strong claims have been made regarding the nature of these "looking at nothing" effects. Altmann (2004, cf. Richardson and Spivey, 2000) has proposed that "the spatial pointers are a component of the episodic trace associated with each item – activating that trace necessarily activates the (experiential) component encoding the location of that item, and it is this component that automatically drives the eyes toward that location" (p. B86). Similarly, Ferreira et al. (2008) claimed that "whether the looks are intentional or are unconsciously triggered, the conclusion is the same: looking at nothing is an entirely expected consequence of human cognitive architecture" (p. 409).

However, Mishra et al. (2011) have found that this is not a universal trait of human cognition. Mishra et al. (2011) studied Indian low literates (2 mean years of formal schooling, but proficient speakers/listeners) and high literates (15 mean years of formal schooling) on the same "look and listen" task as used by Altmann (2004) to test these claims. If "looking at nothing" is an automatic reflex of the cognitive system to refer to previously presented visual objects, then it should be present in all proficient speakers/listeners regardless of their level of formal schooling. High and low literates were presented with a visual display of four objects (a semantic competitor, e.g., "kachuwa," turtle, and three distractors) for 5 s. Then the visual display was replaced with a blank screen and participants listened to simple spoken sentences containing a target word (e.g., "magar," crocodile, a semantic competitor of "kachuwa," turtle). High but not low literates looked at the empty region previously occupied by the semantic competitor as the spoken target word was heard. In a follow-up experiment, the same participants were presented with the identical materials except that the visual display (containing the semantic competitor and the distractors) was present as participants heard the spoken sentences. With such a set up both low literates and high literates did shift their eye gaze toward the semantic competitors immediately as the target word was heard. In another study, Huettig et al. (2011d), found that low literates also made fewer anticipatory eye movements than high literates. Low and high literates (2 and 12 years of schooling) listened to simple spoken sentences containing a target word (e.g., "door") while looking at a visual display of four objects (the target, i.e., the door, and three distractors). The spoken Hindi sentences contained adjectives followed by the (semantically neutral) particle wala/wali and a noun (e.g., "Abhi aap ek uncha wala darwaja dekhnge," Right now you are going to see a high door). Adjective (e.g., uncha/unchi, high) and particle (wala/wali) are gender-marked in Hindi and thus participants could use syntactic information to predict the target. To maximize the likelihood to observe anticipation effects, adjectives which were also semantically and associatively related to the target object were chosen. High literates started to shift their eye gaze to the target object well before target word onset. Low literates' fixations on the targets only started to differ from looks on the unrelated distractors once the spoken target word acoustically unfolded (more than a second later than the high literates).

Further research is currently underway to establish why these populations differ in language-mediated eye movement behavior (see also Huettig et al., 2011c). We know from control tests that they do not depend on IQ. The results are also unlikely to be due to differences in processing 2D information during picture processing. In a recent study we observed very high picture naming accuracy scores in the low literate group. Moreover, in Experiment 2 of Huettig et al. (2011d), low literates were not slower than high literates in their shifts in eye gaze to the target objects *when hearing the target word*, they just did not use contextual information to predict them before the target word was heard. This makes it very unlikely that the observed pattern of results is due to slow information retrieval during picture processing. Instead, we conjecture that literacy is a main factor underlying differences in language-mediated anticipation. To maintain a high reading speed, prediction is helpful if not necessary. Reading and spoken language comprehension, for instance, differ in the amount of information that is processed per time unit (approx. 250 vs. 150 words/min). It has also been observed that readers make use of statistical knowledge in the form of transitional probabilities, i.e., that the occurrence of one word can be predicted from the occurrence of another (McDonald and Shillcock, 2003). Low levels of reading and writing practice greatly decreases the exposure to such word-to-word contingency statistics in low literates. Huettig et al. (2011d) propose that formal literacy may enhance individuals' abilities to generate lexical predictions, abilities that help literates to exploit contextually relevant predictive information in other situations such as when anticipating which object an interlocutor will refer to next in one's visual environment.

In terms of the absence of looks to the semantic competitors by the low literates in the "blank screen" study it is less clear how literacy may have mediated these results. An intriguing possibility is that the well-known "looking at nothing" effects (Spivey and Geng, 2001; Altmann, 2004) reflect merely that participants with high levels of formal education are more familiar with the concept of experimentation and attempt to link "explicitly" the previewed visual display and the unfolding spoken sentence when viewing the blank screen and that low literates are much less likely to do so. A related possibility is that high literates may simply be better in correctly guessing the "purpose" of "blank screen" experiments. Alternatively, it may be that working memory differences underlie the differences between high and low literates' "looking at nothing" behavior. In any case, these results underscore the need to investigate the behavior of non-student participant populations. Ongoing research also examines the attentional basis of these differences between low and high literates. What seems

clear from these data is that the language–vision interaction is modulated by cognitive factors which correlate with formal literacy and/or general schooling and thus accounts which assume that this language-mediated eye movement behavior is automatic or a non-trivial consequence of human cognitive architecture may have to be revised.

FUTURE DIRECTIONS AND CONCLUSION

How will we be most likely to make progress in our understanding of the mechanisms and representations shared by language, vision, attention, and memory during language-mediated eye gaze? Besides a focus on individual and group differences, neuroscientific approaches will undoubtedly prove to be important. For example, activity in different brain areas may reveal at what level linguistic and visual input map onto each other (ranging from occipital to temporal areas), how this is translated into a saccadic signal (ranging from parietal areas to the frontal eye fields, as well as subcortical areas such as the superior colliculus), and to what extent systems are involved that are typically associated with top-down attention and working memory (such as the dorsolateral prefrontal cortex).

Computational modeling will also increasingly play an important role (see Allopenna et al., 1998; Roy and Mukherjee, 2005; Mayberry et al., 2009; Mirman and Magnuson, 2009; Stephen et al., 2009; McMurray et al., 2010; Kukona and Tabor, 2011). An advantage of such models is that theoretical notions and representations underlying language-mediated eye gaze are explicitly exposed. They also allow direct manipulation of representations, processes, and specific factors (e.g., past experience, age of acquisition) which are difficult to control in real participants. In addition, novel predictions about human performance can be derived since models often produce output phenomena which have not been reported previously.

A further fruitful avenue of research is the investigation of brain lesions using single case studies, studies involving groups of patients, or the application of transcranial magnetic stimulation (TMS) on healthy participants. Patients suffering from Balint's syndrome, for instance, have brain damage to the left and right parietal lobes and severe spatial deficits. One particularly interesting symptom is the difficulty that these patients appear to have with the binding of different visual features of an object (e.g., color and shape, cf. Friedman-Hill et al., 1995). One question is whether this type of lesion would also affect the binding of linguistic information to visual locations, as in the visual world paradigm, or whether linguistic information escapes the disintegration that characterizes the visual features.

In sum, we conclude that the investigation of language-mediated eye gaze is a useful approach to study the interaction of linguistic and non-linguistic cognitive processes. The data reviewed suggest that the representational level at which language–vision mapping occurs is co-determined by the timing of cascaded processing in the spoken word and object/visual word recognition systems, by the temporal unfolding of the spoken language, and by the nature of the visual environment (e.g., the characteristics of the visual stimuli, and the possibility of other representational matches). The most concrete proposal regarding attentional mechanisms to date postulates that language-mediated visual orienting arises because linguistic and non-linguistic information and attention are instantiated in the same common coding substrate. We suggest that local microcircuitries involving feedforward and feedback loops may instantiate such a representational substrate. We further conclude that little is currently known about the exact nature of the types of memory involved. Questions that remain to be answered include whether the binding of linguistic types to visual tokens is an implicit or an explicit process, occurs automatically or is subject to cognitive control, whether it is restricted by capacity limitations, and to what extent it suffers from interference and decay. We conjecture that an explicit working memory will be central to explaining interactions between language and visual attention. Though much progress has been made it is clear that a synthesis of further experimental evidence from a variety of fields of inquiry, methods, and distinct participant populations will prove to be crucial for our understanding about how language, vision, attention, and memory interact.

ACKNOWLEDGMENTS

Support was provided by the Max Planck Society (Falk Huettig), a grant from the Department of Science and Technology, India (Ramesh Kumar Mishra), and a VIDI grant from NWO (awarded to Christian N. L. Olivers).

REFERENCES

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439.

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the "blank screen paradigm." *Cognition* 93, 79–87.

Altmann, G. T. M., and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: linking anticipatory (and other) eye movements to linguistic processing. *J. Mem. Lang.* 57, 502–518.

Altmann, G. T. M., and Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cogn. Sci.* 33, 583–609.

Arnett, J. (2008). The neglected 95%: why American psychology needs to become less American. *Am. Psychol.* 63, 602–614.

Bornstein, M. H. (1985). On the development of color naming in young children: data and theory. *Brain Lang.* 26, 72–93.

Cave, K. R. (1999). The feature gate model of visual attention. *Psychol. Res.* 62, 182–194.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cogn. Psychol.* 6, 84–107.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–185.

Cree, G. S., and McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *J. Exp. Psychol. Gen.* 132, 163–201.

Dahan, D., and Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition. *Psychon. Bull. Rev.* 12, 453–459.

Davidoff, J., and Mitchell, D. (1993). The color cognition of children. *Cognition* 48, 121–137.

Dell'Acqua, R., and Grainger, J. (1999). Unconscious semantic priming form pictures. *Cognition* 73, B1–B15.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.

Dunabeitia, J. A., Aviles, A., Afonso, O., Scheepers, C., and Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: evidence from the visual-world paradigm. *Cognition* 110, 284–292.

Ferreira, F., Apel, J., and Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends Cogn. Sci. (Regul. Ed.)* 12, 405–410.

Friedman-Hill, S. R., Robertson, L. C., and Treisman, A. (1995). Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. *Science* 269, 853–855.

Gaskell, M. G., and Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Lang. Cogn. Process.* 12, 613–656.

Hamker, F. H. (2004). A dynamic model of how feature cues guide spatial attention. *Vision Res.* 44, 501–521.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–135.

Hoover, M. A., and Richardson, D. C. (2008). When facts go down the rabbit hole: contrasting features and object hood as indexes to memory. *Cognition* 108, 533–542.

Huettig, F., and Altmann, G. (2011). Looking at anything that is green when hearing "frog": how object surface colour and stored object colour knowledge influence language-mediated overt attention. *Q. J. Exp. Psychol.* 64, 122–145.

Huettig, F., and Altmann, G. T. M. (2004). "The online processing of ambiguous and unambiguous words in context: evidence from head-mounted eye-tracking," in *The Online Study of Sentence Comprehension: Eyetracking, ERP and Beyond*, eds M. Carreiras and C. Clifton (New York, NY: Psychology Press), 187–207.

Huettig, F., and Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition* 96, B23–B32.

Huettig, F., and Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Vis. Cogn.* 15, 985–1018.

Huettig, F., and McQueen, J. M. (2007). The tug of war between phonological, semantic, and shape information in language-mediated visual search. *J. Mem. Lang.* 54, 460–482.

Huettig, F., and McQueen, J. M. (2011). The nature of the visual environment induces implicit biases during language-mediated visual search. *Mem. Cognit.* 39, 1068–1084.

Huettig, F., Olivers, C. N. L., and Hartsuiker, R. J. (2011a). Looking, language, and memory: bridging research from the visual world and visual search paradigms. *Acta Psychol. (Amst.)* 137, 138–150.

Huettig, F., Rommers, J., and Meyer, A. S. (2011b). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychol. (Amst.)* 137, 151–171.

Huettig, F., Singh, N., and Mishra, R. K. (2011c). Language-mediated visual orienting behavior in low and high literates. *Front. Psychol.* 2:285. doi:10.3389/fpsyg.2011.00285

Huettig, F., Singh, N., Singh, S., and Mishra, R. K. (2011d). "Language-mediated prediction is related to reading ability and formal literacy," in *Paper Presented at the AMLaP 2011 Conference in Paris*, Paris.

Huettig, F., Quinlan, P. T., McDonald, S. A., and Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychol. (Amst.)* 121, 65–80.

Humphreys, G. W., and Müller, H. J. (1993). SEarch via recursive rejection (SERR): a connectionist model of visual search. *Cogn. Psychol.* 25, 43–110.

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* New York: Oxford University Press.

Johnson, E., McQueen, J. M., and Huettig, F. (2011). Toddlers' language-mediated visual search: they need not have the words for it. *Q. J. Exp. Psychol.* 64, 1672–1682.

Johnson, E. K., and Huettig, F. (2011). Eye movements during language-mediated visual search reveal a strong link between overt visual attention and lexical processing in 36-month-olds. *Psychol. Res.* 75, 35–42.

Kahneman, D., Treisman, A., and Gibbs, B. (1992). The reviewing of object files: object-specific integration of information. *Cogn. Psychol.* 24, 175–219.

Kanwisher, N. (1987). Repetition blindness: type recognition without token individuation. *Cognition* 27, 117–143.

Knoeferle, P., and Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye-movements. *J. Mem. Lang.* 57, 519–543.

Konkle, T., Brady, T. F., Alvarez, G. A., and Oliva, A. (2010a). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol. Gen.* 139, 558–578.

Konkle, T., Brady, T. F., Alvarez, G. A., and Oliva, A. (2010b). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol. Sci.* 21, 1551–1556.

Kukona, A., and Tabor, W. (2011). Impulse processing: a dynamical systems model of the visual world paradigm. *Cogn. Sci.* 35, 1009–1051.

Lamme, V. A. F., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.* 104, 211–240.

Mani, N., and Plunkett, K. (2010). In the infant's mind's ear: evidence for implicit naming in 18-month-olds. *Psychol. Sci.* 21, 908–913.

Marcus, G. (1998). Rethinking eliminative connectionism. *Cogn. Psychol.* 37, 243–282.

Marcus, G. (2001). *The Algebraic Mind.* Cambridge, MA: MIT Press.

Marslen-Wilson, W., and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* 10, 29–63.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition* 25, 71–102.

Mayberry, M., Crocker, M. W., and Knoeferle, P. (2009). Learning to attend: a connectionist model of the coordinated interplay of utterance, visual context, and world knowledge. *Cogn. Sci.* 33, 449–496.

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86.

McDonald, S. A. (2000). *Environmental Determinants of Lexical Processing Effort.* Unpublished doctoral dissertation, University of Edinburgh, Scotland. Available at: http://www.inf.ed.ac.uk/publications/thesis/online/IP000007.pdf [accessed December 10, 2004].

McDonald, S. A., and Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities. *Psychol. Sci.* 14, 648–652.

McMurray, B., Samelson, V. M., Lee, S. H., and Tomblin, J. B. (2010). Individual differences in online spoken word recognition: implications for SLI. *Cogn. Psychol.* 60, 1–39.

McMurray, B., Tanenhaus, M., and Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86, B33–B42.

Meyer, A. S., Belke, E., Telling, A. L., and Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychon. Bull. Rev.* 14, 710–716.

Meyer, A. S., and Damian, M. F. (2007). Activation of distractor names in the picture-picture interference paradigm. *Mem. Cognit.* 35, 494–503.

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202.

Mirman, D., and Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Mem. Cognit.* 37, 1026–1039.

Mishra, R. K., and Marmolejo-Ramos, F. (2010). On the mental representations originating during the interaction between language and vision. *Cogn. Process.* 11, 295–305.

Mishra, R. K., Singh, N., and Huettig, F. (2011). "'Looking at nothing' is neither automatic nor an inevitable consequence of human cognitive architecture," in *Paper Presented at the AMLaP 2011 Conference in Paris,* Paris.

Moores, E., Laiti, L., and Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nat. Neurosci.* 6, 182–189.

Morsella, E., and Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 555–563.

Moss, H. E., McCormick, S. F., and Tyler, L. K. (1997). The time-course of semantic activation during spoken word recognition. *Lang. Cogn. Process.* 12, 695–731.

Myachykov, A., Thompson, D., Scheepers, C., and Garrod, S. (2011). Visual attention and structural choice in

sentence production across languages. *Lang. Linguist. Compass* 5, 95–107.

Navarette, E., and Costa, A. (2005). Phonological activation of ignored pictures: further evidence for a cascade model of lexical access. *J. Mem. Lang.* 53, 359–377.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). *The University of South Florida Word Association, Rhyme, and Word Fragment Norms.* Available at: http://www.usf.edu/FreeAssociation/

Noizet, G., and Pynte, J. (1976). Implicit labeling and readiness for pronunciation during the perceptual process. *Perception* 5, 217–223.

Olivers, C. N. L. (2009). What drives memory-driven attentional capture? The effects of memory type, display type, and search type. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1275–1291.

Olivers, C. N. L. (2011). Long-term visual associations affect attentional guidance. *Acta Psychol. (Amst.)* 137, 243–247.

Olivers, C. N. L., Meijer, F., and Theeuwes, J. (2006). Feature-based memory-driven attentional capture: visual working memory content affects visual attention. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 1243–1265.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Can. J. Psychol.* 46, 461–488.

O'Regan, J. K., and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–1011.

Palmer, J., Verghese, P., and Pavel, M. (2000). The psychophysics of visual search. *Vision Res.* 40, 1227–1268.

Posner, M. I. (1980). Orienting of attention, the VIIth Sir Frederic Bartlett Lecture. *Q. J. Exp. Psychol.* 32, 3–25.

Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition* 80, 127–158.

Reynolds, J. H., and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron* 24, 19–29.

Richardson, D. C., and Spivey, M. J. (2000). Representation, space and hollywood squares: looking at things that aren't there anymore. *Cognition* 76, 269–295.

Roy, D., and Mukherjee, N. (2005). Towards situated speech understanding: visual context priming of language models. *Comput. Speech Lang.* 19, 227–248.

Salverda, A. P., Brown, M., and Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual-world studies. *Acta Psychol. (Amst.)* 137, 172–180.

Salverda, A. P., Dahan, D., and McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* 90, 51–89.

Salverda, A. P., and Tanenhaus, M. K. (2010). Tracking the time course of orthographic information in spoken-word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 1108–1117.

Schreuder, R., Flores D'Arcais, G. B., and Glazenborg, G. (1984). Effects of perceptual and conceptual similarity in semantic priming. *Psychol. Res.* 45, 339–354.

Shatzman, K. B., and McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychol. Sci.* 17, 372–377.

Soto, D., Heinke, D., Humphreys, G. W., and Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 248–261.

Soto, D., and Humphreys, G. W. (2007). Automatic guidance of visual attention from verbal working memory. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 730–757.

Spivey, M., and Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychol. Res.* 65, 235–241.

Spivey, M. J., Richardson, D. C., and Fitneva, S. A. (2004). "Memory outside of the brain: oculomotor indexes to visual and linguistic information," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. Henderson and F. Ferreira (New York: Psychology Press), 161–189.

Stephen, D. G., Mirman, D., Magnuson, J. S., and Dixon, J. A. (2009). Lévy-like diffusion in eye movements during spoken-language comprehension. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 79, 056114.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634.

Theeuwes, J. (2010), Top-down and bottom-up control of visual selection. *Acta Psychol. (Amst.)* 135, 77–99.

Treisman, A. (1996). The binding problem. *Curr. Opin. Neurobiol.* 6, 171–178.

Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136.

Treisman, A., and Sato, S. (1990). Conjunction search revisited. *J. Exp. Psychol. Hum. Percept. Perform.*16, 459-478.

van der Velde, F., and de Kamps, M. (2001). From knowing what to knowing where: modeling object-based attention with feedback disinhibition of activation. *J. Cogn. Neurosci.* 13, 479–491.

Vanduffel, W., Ekstrom, L. B., Roelfsema, P. R., Arsenault, J. T., and Bonmassar, G. (2008). Bottom-up dependent gating of frontal signals in early visual cortex. *Science* 321, 414–417.

Vickery, T. J., King, L.-W., and Jiang, Y. (2005). Setting up the target template in visual search. *J. Vis.* 5, 81–92.

Wolfe, J. M. (1994). Guided Search 2.0. A revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238.

Wolfe, J. M. (1998). "Visual search," in *Attention,* ed. H. Pashler (Hove: Psychological Press), 14–73.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., and Vasan, N. (2004). How fast can you change your mind? *Vision Res.* 44, 1411–1426.

Wolfe, J. M., Klempen, N., and Dahlen, K. (2000). Post attentive vision. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 693–716.

Yee, E., Huffstetler, S., and Thompson-Schill, S. L. (2011). Function follows form: activation of shape and function features during object identification. *J. Exp. Psychol. Gen.* 140, 348–363.

Yee, E., Overton, E., and Thompson-Schill, S. L. (2009). Looking for meaning: eye movements are sensitive to overlapping semantic features, not association. *Psychon. Bull. Rev.* 16, 869–874.

Yee, E., and Sedivy, J. C. (2001). "Using eye movements to track the spread of semantic activation during spoken word recognition," in *Paper Presented at the 13th Annual CUNY Sentence Processing Conference*, Philadelphia.

Yee, E., and Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1–14.

Zelinsky, G. J., and Murphy, G. L. (2000). Synchronizing visual and language processing: an effect of object name length on eye movements. *Psychol. Sci.* 11, 125–131.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2011; paper pending published: 13 September 2011; accepted: 20 December 2011; published online: 09 January 2012.

Citation: Huettig F, Mishra RK and Olivers CNL (2012) Mechanisms and representations of language-mediated visual attention. Front. Psychology ***2****:394. doi: 10.3389/fpsyg.2011.00394*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 23 December 2011
doi: 10.3389/fpsyg.2011.00376

Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension

Pia Knoeferle*, Maria Nella Carminati, Dato Abashidze and Kai Essig

Cognitive Interaction Technology Excellence Cluster, Bielefeld University, Bielefeld, Germany

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Christoph Scheepers, University of Glasgow, UK
Falk Huettig, Max Planck Institute for Psycholinguistics, Netherlands

****Correspondence:***
Pia Knoeferle, Cognitive Interaction Technology Excellence Cluster, Bielefeld University, Morgenbreede 39, Building H1, D-33615 Bielefeld, Germany.
e-mail: knoeferl@cit-ec.uni-bielefeld.de

Eye-tracking findings suggest people prefer to ground their spoken language comprehension by focusing on recently seen events more than anticipating future events: When the verb in NP1-VERB-ADV-NP2 sentences was referentially ambiguous between a recently depicted and an equally plausible future clipart action, listeners fixated the target of the recent action more often at the verb than the object that hadn't yet been acted upon. We examined whether this inspection preference generalizes to real-world events, and whether it is (vs. isn't) modulated by how often people see recent and future events acted out. In a first eye-tracking study, the experimenter performed an action (e.g., sugaring pancakes), and then a spoken sentence either referred to that action or to an equally plausible future action (e.g., sugaring strawberries). At the verb, people more often inspected the pancakes (the recent target) than the strawberries (the future target), thus replicating the recent-event preference with these real-world actions. Adverb tense, indicating a future versus past event, had no effect on participants' visual attention. In a second study we increased the frequency of future actions such that participants saw 50/50 future and recent actions. During the verb people mostly inspected the recent action target, but subsequently they began to rely on tense, and anticipated the future target more often for future than past tense adverbs. A corpus study showed that the verbs and adverbs indicating past versus future actions were equally frequent, suggesting long-term frequency biases did not cause the recent-event preference. Thus, (a) recent real-world actions can rapidly influence comprehension (as indexed by eye gaze to objects), and (b) people prefer to first inspect a recent action target (vs. an object that will soon be acted upon), even when past and future actions occur with equal frequency. A simple frequency-of-experience account cannot accommodate these findings.

Keywords: visually situated sentence comprehension, eye tracking, visual context effects

1. INTRODUCTION

The role of prediction in language and cognition is a much-debated issue in the cognitive sciences. Prediction plays an important part in accounts of event perception (Zacks et al., 2007), in visual perception (e.g., Nijhawan, 1994; Berry et al., 1999), action anticipation (e.g., Miall et al., 1993; Wolpert et al., 1995; Aglioti et al., 2008), and in theoretical as well as modeling research on language comprehension (e.g., Elman, 1990; Hale, 2003; Federmeier, 2007; Pickering and Garrod, 2007; Levy, 2008). For language comprehension more specifically, the important role of predictive processes is evidenced by both findings from studies recording event-related brain potentials (e.g., Berkum et al., 2005; DeLong et al., 2005) and from studies tracking eye movements (e.g., Altmann and Kamide, 1999; Sedivy et al., 1999; Kamide et al., 2003a,b; see also Aborn et al., 1959; Tulving and Gold, 1963; Fischler and Bloom, 1979, for related early studies on word prediction in sentence context).

In more detail, both the current interpretation and linguistic as well as non-linguistic information from the immediate situation can enable predictive processes during language comprehension. Visual event-related brain potential (ERP) recordings showed that when a definite article (e.g., *an*) was incongruous with the contextually most-expected noun (e.g., *kite* after *The day was breezy so the boy went outside to fly an...*), mean amplitude ERPs to the determiner were more negative going relative to when the determiner was congruous with the contextually most-expected noun (DeLong et al., 2005). Corroborating evidence for predictive processes based on the current utterance interpretation comes from "anticipatory" eye movements to target objects (i.e., eye movements to these objects before they are mentioned). Verb selectional restrictions (Altmann and Kamide, 1999), compositional noun and verb meaning, and associated world knowledge (Kamide et al., 2003a,b), prosody (Weber et al., 2006), or information structure (Kaiser and Trueswell, 2005) can each restrict the range of target objects that can be mentioned next, as evidenced by participants inspecting a target object before its mention relative to a control condition. Anticipatory gaze effects during spoken language comprehension can also be elicited by information from

the immediate non-linguistic context such as the actions that an object affords (Chambers et al., 2004), and verb-mediated depicted events (Knoeferle et al., 2005). In sum, language comprehension is characterized by a forward-looking mechanism that generates expectations about upcoming information based on the current interpretation, related linguistic, and world knowledge, as well as contextual information from the immediate situation.

In addition to information from the immediate situation, *recent* visual context information can also incrementally inform language comprehension (see Altmann, 2004; Knoeferle and Crocker, 2007; Huettig et al., 2011a), and memory task performance (Spivey and Geng, 2001). In Altmann (2004), after participants had inspected a man, a woman, a newspaper, and a cake, the screen went blank. Participants subsequently heard, for instance, *The man will eat*... Shortly after hearing *eat* they inspected the location where they had previously seen the cake before *cake* was mentioned. These findings corroborate the idea that semantic expectations during language comprehension are incrementally related to representations of recently inspected clipart objects (Altmann, 2004). A study by Knoeferle and Crocker (2007) extended these results to quasi-dynamically depicted clipart events and examined how visual interrogation of a scene is informed by information from events that participants had just inspected compared with events they could expect to happen in the near future. Participants saw a character (a waiter) move toward an object, interact with it (e.g., polish candelabra), and move away from it. People subsequently passively listened to an utterance that referred either to the recent action (polishing the candelabra: simple past tense: *Der Kellner polierte kürzlich die Kerzenleuchter*, "The waiter recently polished the candelabra") or to an equally plausible action that hadn't yet been performed (e.g., polishing crystal glasses; present tense with future meaning: *Der Kellner poliert sogleich die Kristallgläser*, "The waiter will soon polish the crystal glasses"). At the verb *poliert*... ("polish..") the comprehension system and visual attention had a choice between anticipating the recent action target versus anticipating (and thus inspecting) the target of the as-yet-unseen future action. Participants preferentially anticipated the target of the recent (vs. the other, future) action, a gaze pattern that continued even as future tense information became available through the adverb (e.g., *sogleich*, "soon"). Verb meaning and future tense information did not elicit expectations of future events and people rather relied on the recently inspected events.

The present paper investigates in more detail how information from recent events compared with expectation of future events affects the visual inspection of (real-world) objects in visual context. Both visual anticipation and processes of accessing visual context information from working memory have been accommodated in existing accounts of situated language comprehension (see the Coordinated Interplay Account, CIA Knoeferle and Crocker, 2007). Overall, the Coordinated Interplay Account is concerned with accommodating the rapid interplay between language comprehension, (visual) attention, and subsequent feedback of non-linguistic visual information into comprehension processes. In line with existing evidence, the CIA assumes that comprehenders incrementally build an interpretation of the sentence and derive associated expectations. The (partial) interpretation built in this first stage directs attention (referentially but also anticipatorily) to relevant aspects of visual context or representations thereof in working memory, and visual context information that is not immediately visually present experiences some decay. The representations of linguistic and non-linguistic content that are in the focus of attention are then co-indexed (e.g., grounding a verb in its action referent), and if necessary the interpretation is revised based on visual context information. As the next word is encountered, this temporally coordinated interplay continues. The three stages can overlap as the sentence is processed but they depend on each other for information.

When considering the observed preference to anticipate the recent (vs. future) event target, the CIA accommodates it via a reference-first mechanism. As people hear the sentence-initial noun "waiter," they mostly inspect the waiter. Then they hear the verb "polish...," and all else being equal they first attempt to ground it in an action (representation) according to the CIA. This leads to participants inspecting the location at which the action took place. Less attention goes toward anticipating the target of future events (at least when a referential competitor – the action – has recently been seen and its target is still present). To the extent that the event representations of the recent events decay, the preference to inspect the recent-event target should decrease.

In the study by Knoeferle and Crocker, 2007, Experiment 3), however, decay was unlikely since for each critical trial only the "recent" event was depicted prior to sentence comprehension (and then referenced in the simple past in the ensuing sentence). The procedure of never depicting the future event may have created a within-experiment frequency bias toward relying more on recently depicted than on equally plausible future events. Perhaps because of this frequency bias, it has been argued that "the fact that the visual world took precedence in these studies over experiential knowledge is not surprising, of course, given that the most reliable cue to who is doing what to whom is whoever one sees doing it, not whoever one thinks is doing it. [...] no input is more privileged than another except insofar as one may be more predictive than the other in a given situation" (see, e.g., Altmann and Mirković, 2009, p. 596f).

These statements were made in the context of an alternative account of situated language processing by Altmann and Mirković (2009). In their model, information from the linguistic and non-linguistic visual context appears as representationally equivalent and to the extent that these two information sources are equally predictive none of them is preferred in predicting what will be mentioned next. Attention is allocated to objects through overlap between object representations (which are assumed to encode an object's location), and linguistic representations derived from the unfolding utterance. The crux in interpreting the Altmann and Mircović account and their statements about the findings from Knoeferle and Crocker, 2007, Experiment 3 lies in understanding what these authors mean by a "reliable" cue and by input being only privileged to the extent that it is more "predictive." The precise meaning of these terms in their paper isn't explicitly defined, rendering their interpretation somewhat problematic. We believe that one "strong" but logically coherent interpretation of these terms within their account is that short-term and/or long-term experience of a given cue determines its predictiveness of subsequent input within their account. This is plausible since experience-based

knowledge and learning also play an important role in the Altmann and Mirković account which views language processing as governed by a mechanism that "[. . .] learns to anticipate, on the basis of its current and preceding input, what input may follow" (Altmann and Mirković, 2009, p. 589). Indeed, learning of statistical regularities is a hallmark of the connectionist network that Altmann and Mirković refer us to in illustrating their account (Altmann and Dienes, 1999). Thus, in the absence of a clear definition of the reliability and predictiveness of a cue we instantiated predictiveness as the short-term frequency with which a participant experienced recent versus future events and long-term regularities of temporal cues in the sentences (e.g., past vs. future tense adverbs).

There are other considerations as to why a frequency-based account of Knoeferle and Crocker's (2007, Experiment 3) findings is not implausible. In fact, in recent years it has become increasingly clear that human language comprehension and also other cognitive and motor processes are exquisitely sensitive to statistical regularities. In action execution, the recent trial-to-trial visuomotor experience can affect upcoming movement decisions (e.g., which one of two potential targets to reach for, Chapman et al., 2010). In language acquisition, statistical regularities can be exploited by children as young as 8 months for segmenting words in fluent speech (Saffran et al., 1996). Short-term linguistic experience can also modulate language production (Kaschak et al., 2006; Haskell et al., 2010) and sentence reading (Wells et al., 2009). Systematic co-variation (vs. random pairing) of novel target and distractor objects speeded up response latencies in identifying the target in a visual search task, suggesting that participants learned the associations between these two objects (Chun and Jiang, 1999).

Overall, then short-term experience of statistical regularities appears to play an important role in a number of cognitive and motor processes. To the extent that the importance of statistical regularities extends to perceptual experience of events, the frequency with which events are shown and then mentioned ("recent events") versus the frequency with which events are performed after they were announced ("future events"), could plausibly affect how rapidly comprehenders access those events, and which ones they prefer to attend to during comprehension. An account in terms of short-term event experience could accommodate the rapid and preferred reliance on recent events when people only see recent and never future events (i.e., a bias of 100:0 toward recent events as was the case in Experiment 3 by Knoeferle and Crocker, 2007). Importantly, a short-term frequency account of the preferred reliance on recent events would further predict that as the ratio of recent versus future events that people perceive reaches a 50:50 frequency distribution (and assuming there is no linguistic frequency bias), the preferred inspection of the target of the recent event should be eliminated.

Alternatively (or in addition), the observed gaze pattern could be caused by comprehenders' long-term linguistic experience. The recent actions in the waiter-polishing study were referred to by a verb in the simple past and an ensuing past tense adverb. The future events were indicated by a verb in the present tense with a future meaning and an ensuing future tense adverb. To the extent that the past tense verbs and adverbs may be more frequent than the present tense verbs and future adverbs, they might be processed more rapidly, cueing comprehenders to preferentially inspect the target of the (recent) action that they refer to (see Dahan et al., 2001) for evidence on the effects of lexical frequency on visual attention to objects during spoken language comprehension.

In the original study (Knoeferle and Crocker, 2007, Experiment 3), people only ever saw the recent (and never the future) event on each trial. Seeing a recent event was thus a cue that was reliably followed by mention of the recent event. In contrast, hearing a sentence about a future event was never followed by actually seeing that future event, and thus an unreliable cue about future events. With a 50:50 frequency distribution, participants see an event and then it's mentioned for half of the critical trials while on the other half of the trials, they hear an event mentioned and then they see it performed. According to the strong version of Altmann and Mircović's (2009) account (see above), recent events should arguably be no more predictive than future events, and so a short-term frequency account would predict no particular reliance on recent actions.

Note that this is only a test of the Altmann and Mircović account to the extent that their account instantiates cue predictiveness exclusively via such statistical regularities. One could argue within their account that seeing an action increases the activation of that action representation. The action representation then overlaps with the representation of a corresponding verb, and modulates the attentional state such that the probability of an eye movement to the location associated with the activated action representation increases. In this way, the account might appear to predict more inspections to the target of the recent (vs. future) action. We think, however, that this argumentation logic doesn't hold for a 50:50 within-experiment frequency distribution of recent versus future events. In the latter case, both remembering a recent action and anticipating a future action is equally predictive of what is mentioned/happens next. Thus, it would appear plausible that even after perception of one action activates its representations, the representations of other, relevant future events is activated just as much upon encountering the verb. Verb overlap with the future event representation could then boost the activation of those representations and modulate the attentional state such that the probability of saccades to the target of a plausible future event increases.

The Coordinated Interplay Account, in contrast, because of its mechanism of first grounding a referent would predict that even with a 50:50 frequency distribution of recent to future events, people should prefer to ground the verb in the recent action and its associated target. Thus, implementing a 50:50 frequency distribution of recent relative to future events (and controlling for linguistic frequency biases of the verbs and adverbs) would permit us to tease apart predictions of a reference-first mechanism from an account that rather emphasizes the predictive nature of an information source as instantiated by short-term frequency of event experience.

Two eye-tracking experiments and a corpus study addressed this question. To ensure that findings generalize to real-world environments, the present studies relied on real-world actions performed by the experimenter (see **Figure 1**). Experiment 1 aimed to replicate the findings from Experiment 3 in Knoeferle and Crocker (2007) with real-world action events, i.e., participants only ever

FIGURE 1 | Example video snapshot: the experimenter sugars the pancakes.

saw an event prior to sentence comprehension on each trial. The subsequent sentence either referred to that event (in the simple past) or it referred to another equally plausible event that could happen in the future. There was thus a 100:0 within-experiment frequency bias toward seeing recent (vs. future) events. By contrast for critical trials in Experiment 2, participants saw the experimenter perform one action prior to the sentence, and the other (future) action after sentence comprehension and overall in that study, the frequency distribution of recent relative to future actions was 50:50. Both the CIA and a short-term frequency instantiation of the account by Altmann and Mirković would predict a recent action preference in Experiment 1. In contrast, for Experiment 2, the CIA (but not a short-term frequency account) would appear to predict a preference to anticipate the target of the recently inspected event. An additional corpus study was conducted to gain insight into whether there was any linguistic bias such that past tense verbs and adverbs might be more frequent than present tense verbs and adverbs indicating future actions.

2. MATERIALS AND METHODS

2.1. PARTICIPANTS

Twenty-four German native speakers (aged 19 to 33, M = 24.83; 8 males, 16 females) participated in Experiment 1, and a further twenty-four native German speakers participated in Experiment 2 (aged 19 to 33; M = 24.92, 12 males, 12 females). Participants (all students of Bielefeld University, Germany) were each paid 4 Euros to take part in the experiments. They all had normal or corrected-to-normal vision, were unaware of the purpose of the experiment and all gave informed consent in accordance with the Declaration of Helsinki.

2.2. MATERIALS AND DESIGN

We created twelve experimental items that each consisted of two everyday objects (e.g., strawberries and pancakes) and four sentences, recorded by a male native German speaker (see **Table 1** for an example). Critical sentences were about the two objects and grouped into two tense conditions (future: 1a and recent: 1b). In the future condition, a present tense verb with a temporal adverb (*demnächst*, "soon") indicated the future. In the recent condition, tense was marked on the last letter of each verb (e.g., the *-e* in *zuckerte*, "sugared") and via the temporal adverb (*kürzlich*, "recently"). For the experimental sentences all words were matched for spoken syllables and lemma frequency within an item (Baayen et al., 1995). The counterbalancing versions (1a′ and 1b′ for 1a and 1b respectively) served to present each object once as the target of a recent, and once as the target of a future action, ensuring that visual characteristics of a post-verbal target object contributed equally to each of the two conditions.

Table 1 | Example item set for Experiments 1 and 2.

1a	Future condition	*Der Versuchsleiter zuckert demnächst die Erdbeeren.* "The experimenter will soon sugar the strawberries."
1a′	Future condition	*Der Versuchsleiter zuckert demnächst die Pfannkuchen.* "The experimenter will soon sugar the pancakes."
1b	Recent condition	*Der Versuchsleiter zuckerte kürzlich die Pfannkuchen.* "The experimenter recently sugared the pancakes."
1b′	Recent condition	*Der Versuchsleiter zuckerte kürzlich die Erdbeeren.* "The experimenter recently sugared the strawberries."

1a and 1b are examples of the conditions; 1a′ and 1b′ are the corresponding counterbalancing sentences.

The two objects of each experimental trial (e.g., strawberries and pancakes) could undergo the same action (e.g., sugaring). Experimental sentences about these objects began with *Der Versuchsleiter* ("The experimenter") followed by the verb (e.g., *zuckert*. . ., "sugar. . ."). Because of the counterbalancing, the two objects were equiprobable as targets of the action. Prior to the end of the verb (e.g., *zuckert*. . ., "sugar. . .") sentence tense was ambiguous and the sentence could thus either refer to a recent event (e.g., the experimenter had just sugared the pancakes) or to a future event (e.g., sugaring the strawberries). As the verb ending (*-e* in *zuckerte*, "sugared") and the adverbs were encountered, people could rely on the temporal cues to anticipate the recent versus the future event, although we know that prior research has reported weak effects of tense (Knoeferle and Crocker, 2007). However, the sentence-final noun phrase refers to the target of the recent (1b) versus future (1a) event; so, soon after people start processing this noun phrase, we should begin to see more eye gaze to the correct target (recent condition: pancakes, 1b; future condition: strawberries, 1a).

In Experiment 1, the experimenter performed only one action before the sentence for each experimental item (e.g., sugaring the pancakes), and then participants either heard a spoken sentence in the past (1b, **Table 1**) or in the future (1a, **Table 1**) condition. Participants thus saw 100:0 recent (vs. future) events and heard an equal number of sentences in the recent and future condition. In Experiment 2, the experimenter performed one action before the sentence (sugaring the pancakes), and another action after sentence presentation (sugaring the strawberries) on each critical trial such that participants not only heard equally many recent and future event sentences but also saw 50:50 recent to future events.

In addition to the twelve experimental items we created 24 filler sentences. These ensured that participants were exposed to a range of sentence and action combinations. Filler sentences were identical in the two experiments. They contained a verb in the past tense on 12 trials, and a verb in the present tense for the other 12 trials. In 8 filler trials the adverb indicated the recent past (4 trials) or the near future (4 trials). Adverbs for the other 16 filler sentences did not indicate a point in time but expressed mood, or degree of certainty of an event. The filler trials differed between the experiments in when people saw an action. In Experiment 1, the experimenter performed one action on each trial, prior to sentence presentation. In Experiment 2, for 8 of the filler trials, the experimenter conducted the action as the sentence was spoken. For another 8 filler trials, people only saw one action before sentence presentation (4 trials), or one action after sentence presentation (4 trials). For a further 8 filler trials, participants saw an action both before and after the sentence was presented. From the sentences in the two conditions and their two counterbalancing versions we created four lists using a Latin square. Each list contained every item in only one condition and all 24 filler sentences. Lists were pseudo-randomized and each participant saw an individually randomized version of one of the four experimental lists.

2.3. PROCEDURE

Participant were seated opposite the experimenter in front of a table. They were informed that the experiment would use an eye tracker (SMI iView X HED mobile), and they were calibrated using a 5-point calibration routine. When calibration was successful, the experiment started. Prior to the experiment, participants were instructed to look carefully at the items on the table and listen attentively to the recording played through the loudspeakers. There was no other task. For each trial, the experimenter first put the necessary objects (such as strawberries and pancakes) on the table. For the critical trials in Experiment 1, the experimenter then put sugar on the pancakes (ca. 1500 ms) and subsequently participants listened to German versions of "The experimenter sugars soon the strawberries" or "The experimenter sugared recently the pancakes" (1a and 1b, see **Table 1**). For the critical trials in Experiment 2, the experimenter performed a further action (e.g., sugaring the strawberries), after sentence presentation such that people always saw one action before, and one action after sentence presentation for the critical trials. The experiment lasted approximately 30 min. After the experiment, participants were debriefed.

2.4. ANALYSIS

2.4.1. Eye-tracking data

For the coding of participants' eye gaze during the experimental trials, a period of interest was defined, starting from the onset of the verb until the offset of the post-verbal NP (NP2). The onsets of the critical words in the sentence (verb, adverb, NP2) were marked in the video files using the annotation software ELAN (a tool developed at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, and downloadable at http://www.lat-mpi.eu/tools/elan, see also Sloetjes and Wittenburg, 2008). In the videos, participants' gaze during the trial appeared as a red circle. The duration of each frame in the videos was 40 ms. For the period of interest, participants' fixations were manually coded frame-by-frame as to which region of the scene was fixated in that particular frame. Three regions were defined: the recent target object, the future target object and "other" (i.e., other parts of the scene, for example, the experimenter or the background).

The measure of interest for the purpose of our study is fixations to the recent and future target objects as the sentence unfolds. Using the frame-by-frame gaze data, we first computed gaze probabilities to the two targets in each of the 40 ms time frames. Because looks to these two entities are not linearly independent (more looks to one object imply fewer looks to the other, and vice-versa), we next computed mean log gaze probability ratios for the recent relative to the future target ln (P (recent target)/P (future target)). This measure, which expresses the bias of inspecting the recent relative to the future target, does not violate the linear independence assumption (e.g., Arai et al., 2007; Carminati et al., 2008). In this measure, a score of zero indicates that both targets are fixated equally frequently; a positive score reflects a preference for looking at the recent target over the future target, and a negative ratio indicates the opposite.

For the inferential analyses we defined the following three time windows: the verb region (from verb onset until adverb onset, $M = 1148$ ms); the adverb region (from adverb onset until the offset of the adverb, $M = 1332$ ms) and the NP2 region (from NP2 onset until NP2 offset, $M = 710$ ms). We aggregated mean log gaze probabilities ratios ln (P (recent target)/P (future target)) over each of the three time regions of interest. A further advantage of using log-ratios (in addition to the independence assumption) is that they yield data distributions that are more suitable for parametric testing (standard probabilities often imply a violation of the homogeneity of variance assumption because they have a limited range from 0 to 1; in contrast, log-ratios can take values between minus infinite and plus infinite, which is what is required for parametric testing).

We fitted linear mixed effect (LME) models to the log probability ratios for each of the time regions, using the R-software (version 2.2.0; CRAN project; R Development Core Team, 2008)[1]. Separate models were fitted on log-ratios averaged over participants and items respectively (Barr, 2008). In all models, the predicted outcome was the log ratio of fixations to the recent target relative to the future target and the fixed effect predictor was condition (future vs. recent). To minimize collinearity, we used effect coding by transforming the fixed effect into a numerical value and centering it so as to have a mean of zero and a range of 1 (Baayen, 2008). Effect coding has the further advantage of allowing the coefficients of the regression to be interpreted as the main effects in a standard ANOVA (Barr, 2008). Furthermore, with this coding the intercept represents the estimate of the grand mean; therefore, applied to our particular data, a significant intercept would indicate that the mean log gaze probability ratio ln (P (recent target)/P (future target)) is significantly different from zero. In turn, this would indicate that there is a significant bias toward

[1] Due to a sparse frequencies in the design table we could not rely on the hierarchical log-linear analyses (Scheepers, 2003; Knoeferle and Crocker, 2007).

looking at one object relative to the other, whether or not a significant effect of condition is also present (recall that a log ratio of zero would indicate that there is no such bias). For each analysis, two models were fitted, one including only the random intercept (i.e., allowing the intercept to vary across participants and items respectively) and another including both the random intercept and the random slope (i.e., allowing also the slope of the fixed predictor to vary across the random variables). These models were then evaluated using a log-likelihood ratio test (Baayen, 2008, p. 276) and the more complex model was retained only if it fitted the data significantly better than the simpler one (indicated in **Table 3** with §). A coefficient was considered to be significant at *alpha* = 0.05 when the absolute value of t was greater than 2 (Baayen, 2008)[2].

2.4.2. Corpus data

For the corpus study we looked at five different corpora: the Europa Parliament Corpus (Koehn, 2005), the German Reference Corpus (COSMAS II, Kupietz et al., 2010), deWac (Baroni et al., 2009), Google, and DLex (http://www.dlexdb.de; Heister et al., 2011). We report two different analyses. (1) For our recent condition, we searched for the exact verb forms in the simple past and present perfect to get an estimate of how often people encounter a verb form referring to the past; for the future condition we searched for verb forms in the present tense. (2) We did a frequency count of the temporal adverbs in the two conditions (recent condition: *soeben*, "just now"; *unlängst*, "not long since"; *kürzlich*, *vorhin*, "a little while ago"; future condition: *sogleich*, "presently"; *nachher*, "subsequently"; *demnächst*, "soon"; *baldigst*, "as soon as possible"). A third analysis in which we searched for the exact verb and adverb sequences of our items had to be abandoned due to data sparseness. We obtained frequencies of the verbs and adverbs for each item and normalized these frequencies for each corpus using the number of words in the respective corpus[3]. Since this resulted in small numbers, we multiplied each thus-obtained frequency by 1,000,000 to facilitate interpretation. We present descriptive frequencies of the verb forms and of the adverbs averaged across the individual items (**Table 4**). To ascertain whether there were reliable differences in the frequency scores for our items across the five corpora, we computed the average frequency scores across the five corpora by items (i.e., the 12 verbs used in our study and the 4 temporal adverbs for each condition). We provide the 95 percent confidence interval of the average difference scores for our verbs and adverbs in each of the two conditions (past minus present/future condition for the normalized, multiplied, and averaged scores, **Table 4**).

3. RESULTS

We first present the results of the two eye-tracking studies and subsequently the results of the corpus study. For the eye-tracking data, **Figures 2** and **3** plot the mean log gaze probability ratios computed using the original 40 ms frame data, for the period from verb onset to NP2 offset, for Experiments 1 and 2 respectively. Descriptively, these two graphs reveal an overall preference for looking at the recent target relative to the future target throughout the verb and adverb, shown by the fact that during most of this period the log ratio remains well above zero (indicating that the recent target receives more looks than the future target). As participants hear the second noun, they begin to shift gaze to the future target (the referent of "strawberries") in the future more than in the recent

[2]In choosing to run LME models on data aggregated up to the participant and item level separately, we follow the second approach outlined in Barr (2008) for analyzing visual-world eye-tracking data. It should be noted that this approach is essentially equivalent to running separate repeated measures mixed-design ANOVAs with participants and items as random effects.

[3]The only exception was the Google corpus for which we set the size to 1 since its exact size was unknown.

FIGURE 2 | Mean log gaze probability ratios ln (P(recent target)/P(future target)) as a function of condition from Verb Onset for Experiment 1.

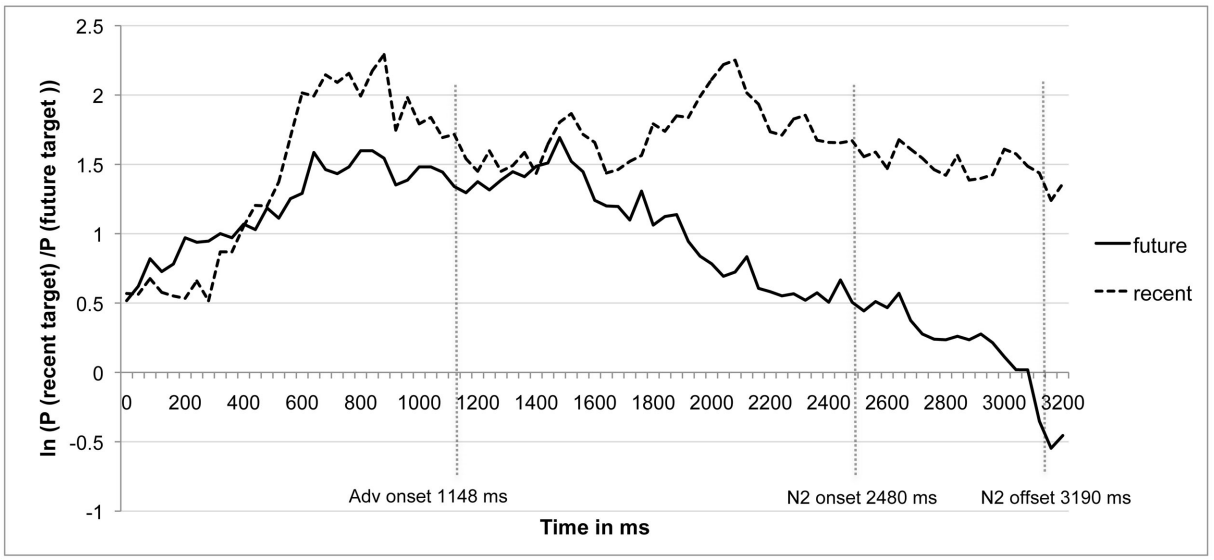


FIGURE 3 | Mean log gaze probability ratios ln (P(recent target)/P(future target)) as a function of condition from Verb Onset for Experiment 2.

condition; the recent target (the referent of "pancakes") is fixated more in the recent condition than in the future condition.

This descriptive pattern was corroborated by the per-region descriptive (**Table 2**) and inferential (**Table 3**) analyses. **Table 2** shows the mean log gaze probability ratios (participants' means) for the three time regions of interest as a function of condition, and **Table 3** summarizes the results of the corresponding LME analyses. As one can see from the means in **Table 2**, there is a general inspection bias in favor of the recent target over the future target object, which we noted in the time course graphs (see **Figures 2** and **3**). Statistical analyses confirmed that this bias was reliable across all three regions in both experiments (i.e., the intercept was significantly different from zero). The positive coefficient for the intercept in **Table 3** indicates that people look more at the recent than future target throughout the sentence.

In the verb region for both experiments, the visual preference for the recent target was not modulated by whether the verb was in the present (future condition) or in the past (recent condition), as evidenced by the absence of a significant main effect of condition. However, as participants incrementally processed the remainder of the sentence, the main effect of tense (future vs. recent) becomes reliable, in the final NP2 region in Experiment 1, and in the adverb and NP2 regions in Experiment 2. In the recent condition, this effect is driven by an increase in the log ratio (i.e., looks to the recent target increase and those to the future target decrease), while in the future condition there is a corresponding decrease in the log ratio (i.e., looks to the recent target decrease and those to the future target increase, see **Table 2**). Concurrent analyses on the same log-ratio measures using mixed-design ANOVAs with participants and items as random effects yielded results in agreement with the LME analyses (see footnote 2).

In the above LME analyses the (positive) grand mean intercept was significantly different from zero, indicating a visual bias toward the recent over the future target averaged over the two

Table 2 | Mean log gaze probability ratios *ln* (*P* (recent target)/*P* (future target)) as a function of condition and time region for Experiment 1 and 2.

Time region	Future condition (present tense verb and future adverb)	Recent condition (past tense verb and adverb)
EXPERIMENT 1		
Verb	1.52 (0.23)	1.51 (0.23)
Adverb	1.16 (0.26)	1.64 (0.36)
NP2	0.35 (0.26)	1.78 (0.28)
EXPERIMENT 2		
Verb	1.43 (0.26)	1.58 (0.22)
Adverb	1.38 (0.25)	2.23 (0.28)
NP2	0.15 (0.20)	2.10 (0.35)

Standard errors are in parentheses.

conditions. To determine the extent to which this visual bias is present in the two separate conditions, particularly in the future condition, we conducted one-sample two-tailed t-tests on the log-ratios of participants and items respectively. These tests, adjusted for two comparisons using the Bonferroni method (new *alpha* level: $0.05/2 = 0.025$), were aimed at ascertaining whether the log-ratio means for each condition are significantly different from zero. With regard to the future condition in Experiment 1, the t-tests were significant in both the verb and adverb region (all $ps < 0.001$), but not in the NP2 region ($p1 = 0.19$, $p2 = 0.16$). This pattern of results was replicated for the future condition of Experiment 2 (verb and adverb region all $ps < 0.001$; NP2 region: $p1 = 0.47$, $p2 = 0.79$), suggesting that the 50/50 manipulation of Experiment 2 was not able to override the visual preference for the recent object found in Experiment 1 in the verb and adverb region. As expected, the t-tests in the recent-event condition achieved

significance for all of the analysis regions in both Experiment 1 and 2 (all $ps < 0.001$).

Table 4 shows the results from the corpus study. It displays the normalized verb and adverb frequencies for the future compared with recent condition. The difference scores (past minus present tense) illustrate that present tense verb forms are descriptively somewhat more frequent than past tense verbs in four (European Parliament, Cosmas II, deWac, and Google) out of five of the analyzed corpora. The table also presents the normalized frequencies for the adverbs which show that the future tense adverbs are more frequent than the past tense adverbs in three (deWac, Google, and DLex) of the five analyzed corpora.

These descriptive trends, however, were not confirmed by the confidence intervals for the difference scores (past minus present tense verbs/adverbs). With the exception of the European Parliament corpus for the adverb counts, the confidence intervals for all of the corpora contained zero, suggesting that the underlying means do not differ reliably. Overall thus, past tense verbs and adverbs in our sentence stimuli do not appear to be more frequent than present tense verb forms and adverbs indicating the near future.

Table 3 | Linear mixed effect model results for Experiments 1 and 2 by time region.

Time region	Coefficient participants	Items	*t*-Value participants	Items
EXPERIMENT 1				
Verb				
Intercept	1.51	1.14	9.20*	8.00*
Cond	−0.02	−0.11	−0.15	−0.78
Adverb				
Intercept	1.40	1.04	5.54*	11.67*
Cond	0.24	0.10	1.31	1.19
NP2				
Intercept	1.06	0.94§	5.14*	5.83§*
Cond	0.72	0.50	4.11*	3.01§*
EXPERIMENT 2				
Verb				
Intercept	1.50	1.30	7.69*	12.01*
Cond	0.07	0.07	0.52	0.98
Adverb				
Intercept	1.80	1.55§	8.30*	7.53§*
Cond	0.43	0.53§	2.74*	2.65§*
NP2				
Intercept	1.12§	0.98§	4.95§*	4.54§*
Cond	0.98§	0.94§	5.56§*	3.83§*

**The effect is significant at alpha = 0.05 (using the $|t| > 2$ criterion).*

§These values refer to the model that has both random intercepts and random slopes; all other values are in respect of models with only random intercepts.

4. DISCUSSION

Two eye-tracking studies assessed whether the frequency with which participants saw recent (vs. future) everyday events within the experiment can eliminate a previously observed preference to inspect recent-event targets more than future event targets after hearing a sentence beginning that was compatible with either event. In NP1-V-ADV-NP2 sentences the verb was referentially ambiguous between a recent action (and its associated target) and an equally plausible future action (and its different target object). When participants saw the experimenter perform only one action per trial, prior to presentation of the spoken sentence (Experiment 1), they more often inspected the target of that recent action than the target of the future event during and shortly after the verb. This confirmed that the time course and qualitative gaze pattern from a clipart eye-tracking experiment (Knoeferle and Crocker, 2007, Experiment 3) extend to real-world actions. The recent-event preference persisted even when participants saw the experimenter perform equally many actions prior to

Table 4 | Normalized frequency counts for the verb forms and adverbs in our materials averaged across the items.

	European Parliament (25–30M)	Cosmas II (2000M)	deWac (1411M)	Google set to 1	DLex (100M)
Past tense verb forms	0.091	3.787	0.032	37175.20	8.034
Present tense verb forms	1.123	5.439	0.034	105627.80	3.498
Verb difference scores	−1.032	−1.652	−0.002	−68452.6	4.536
lower/upper 95 CI of the difference scores	−2.779/0.715	−5.843/2.541	−0.0171/0.0130	−247001.4/110096.2	−1.750/10.822
Adverbs indicating the past	58.879	27.774	0.183	417298.0	18.680
Adverbs indicating the future	11.537	17.012	0.184	421338.4	23.805
Adverb difference scores	47.343	10.762	−0.001	−4040.404	−5.125
lower/upper 95 CI of the difference scores	1.126/93.559	−20.667/42.191	−0.239/0.237	−528146.7/520065.9	−45.065/34.815

"Past tense verb forms" and "present tense verb forms" indicate the averaged and normalized frequencies for the recent and future conditions respectively. "Adverbs indicating the past" and "adverbs indicating the future" present the averaged and normalized frequency averages across the adverbs used in the recent and future conditions. "Verb difference scores" and "Adverb difference scores" present the results for subtracting the scores for verbs/adverbs in the future from those for verbs/adverbs in the recent condition. Negative difference scores indicate lower frequencies for the past than present tense verbs and the adverbs. For each corpus we show the number of tokens in millions (M) in brackets. For the verb and adverb difference scores we list first the lower and then the upper 95 percent confidence interval.

versus after sentence presentation (i.e., recent versus future actions respectively) in Experiment 2.

Overall, the data provide good evidence that people prefer to ground their expectations and visual attention during incremental language understanding more through directing their attention at the target of a recent event than at the target of another, equally plausible, future event. We examined this recent-event preference under two frequency distributions of recent relative to future events (i.e., when there was a frequency bias toward recent events in Experiment 1 and when recent and future events occurred equally often in Experiment 2). Together, these two frequency manipulations permit us to tease apart two competing accounts of how contextual information is used to inform expectations during language comprehension: while both the Coordinated Interplay Account (CIA, Knoeferle and Crocker, 2007) and a short-term frequency instantiation of cue reliability in the account by Altmann and Mirković would have predicted a reliance on recent events time-locked to the verb in Experiment 1, their predictions differ for Experiment 2. Consider their predictions for Experiment 1: The CIA incorporates a reference-first mechanism such that comprehenders upon interpreting a word and all else being equal, first look to ground it and find an appropriate referent. Upon hearing a verb, people should thus engage in a search for a suitable referent (visually by interrogating the scene, or by focusing attention on relevant representations in working memory). A short-term frequency instantiation of the account by Altmann and Mirković also predicts a rapid and preferred reliance on recent depicted events in Experiment 1 but for a different reason – because these events are more predictive of what will be mentioned next (as instantiated via a 100:0 frequency bias toward recent events).

When people saw a 50:50 distribution of recent versus future events in Experiment 2, the predictions made by these two accounts diverge. The CIA would still predict a recent-event preference based on its reference-first mechanism. In contrast, a short-term frequency instantiation of cue reliability would no longer predict a preference to inspect the recent-event target more than the future event target since neither of these two information sources is more predictive of which object will be mentioned next or of which action the verb refers to. Both events and verb/adverb forms are equally frequent within the experiment. Thus having seen one action, the ensuing sentence could 50:50 refer to that recent action vs. an equally plausible future action. The findings from Experiment 2 thus provide support against a purely frequency-based account of cue predictiveness in visually situated utterance comprehension. Apparently short-term, within-experiment perceptual and communicative experience that could immediately have informed comprehender's expectations, did not eliminate the preference to inspect the recent-event target during language comprehension.

As mentioned in the introduction, an alternative possibility is that the past tense verbs and adverbs that we used may be more frequent in long-term experience than their present tense counterparts, and that such a long-term frequency bias could guide visual attention to objects. If such a bias exists we may assume that it can rapidly guide attention, since we know that long-term word frequency has rapid effects on language processing and visual attention in comprehension tasks during reading (e.g., Rayner and Raney, 1996), as well as during spoken language comprehension in visual contexts. For the latter situation, Dahan et al. (2001) found that people fixated objects with frequent (vs. relatively more infrequent) names faster. However, we can be relatively certain that the recent events preference indexed via visual attention that we observed in both experiments during the verb is not driven by the long-term frequency of occurrence of these words since there was no reliable frequency difference between verbs and adverbs in four out of five examined corpora.

The absence of immediate short-term frequency effects is somewhat surprising in light of existing evidence showing that short-term frequencies can affect a range of cognitive processes, among them action execution (e.g., Chapman et al., 2010), language acquisition (e.g., Saffran et al., 1996; Saffran, 2003), language production (Kaschak et al., 2006; Haskell et al., 2010), sentence reading (Wells et al., 2009), and visual perception (e.g., Chun and Jiang, 1999). And yet, participants in the present experiments were not immediately (during the verb) sensitive to the within-experiment frequency distribution of the recent compared with the future event. Had they been immediately sensitive, we should have seen no difference in target object inspection during the verb in Experiment 2. This is not to say that the 50:50 frequency manipulation in Experiment 2 (relative to Experiment 1) did not modulate visual attention. Indeed, effects of tense in Experiment 2 occurred earlier (during the adverb) than in Experiment 1 (during the post-adverb noun phrase). This confirms that our frequency manipulation was effective, in line with previously observed effects of short-term experience on cognitive processes. A short-term frequency account can accommodate the earlier tense effects in Experiment 2 compared with Experiment 1. By contrast, it cannot accommodate the visual preference for the recent-event target during the verb and adverb in the future condition in Experiment 2.

The Coordinated Interplay Account accommodates this latter gaze pattern by postulating that people prefer to first ground the verb in the recent action, and in the absence of the action they do so by inspecting the target object upon which they had previously seen the action performed. Another (speculative) possibility is that the order in which we experience events and hear them talked about affects our reliance on them during comprehension. Seeing an event and hearing it subsequently talked about as part of our experience, may anchor that event in a different way in our (working) memory compared to predicting an event that then happens. To the extent that this holds, the reported findings contribute toward delineating the role of expectation-based processes in language and cognition. They fit well with other findings that have shown older (vs. younger) adults engage less in predictive processing (Federmeier et al., 2002; Federmeier, 2007), as do high (vs. low) literates (e.g., Huettig et al., 2011b). In the present task, participants were instructed to pay attention to both the visual context and to language. When people had seen an action, they likely kept that action in their working memory. It is possible that working memory representations of the recently seen action increased visual attention to associated objects. Such a view would appear compatible with findings that suggest visual orienting can be guided by the contents of working memory in memory tasks (e.g., Spivey and Geng, 2001), and in

visual search tasks (even when they are not relevant for the ongoing search task, e.g., Olivers et al., 2006). To the extent that these findings extend to language paradigms, they underscore the role of working memory representations in language processing (see also Altmann, 2004; Knoeferle and Crocker, 2007; Huettig et al., 2011a).

This position is compatible with the Coordinated Interplay Account to the extent that the verb representation mediates the retrieval of working memory representations of an action. The result of verb-mediated referential processes is that (visual) attention goes preferentially to the location and target associated with a recent action (vs. anticipating the target of a future event). Future studies will examine role of working memory in the present findings by further increasing the frequency of future events in the experiment and by means of a post-experiment memory test on the recent versus future actions. While it's not entirely clear yet why we observed the recent-event preference in the absence of frequency biases, it is clear that simple, short-term event experience cannot accommodate these findings.

ACKNOWLEDGMENTS

This research was supported by the Cognitive Interaction Technology Excellence Cluster (funded by the German Research Council) and the CRC 673 "Alignment in Communication," both at Bielefeld University, Germany. We thank Eva Mende and Linda Krull for assistance with coding of the video data, Patrick Bremehr for recording the sound stimuli, and Christian Pietsch for advice on the corpus analyses.

REFERENCES

Aborn, M., Rubenstein, H., and Sterling, T. D. (1959). Sources of contextual constraint upon words in sentences. *J. Exp. Psychol.* 57, 171–180.

Aglioti, S. M., Cesari, P., Romani, M., and Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nat. Neurosci.* 11, 1109–1116.

Altmann, G. T. M. (2004). Language-mediated eye-movements in the absence of a visual world: the "blank screen paradigm". *Cognition* 93, B79–B87.

Altmann, G. T. M., and Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science* 284, 875.

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264.

Altmann, G. T. M., and Mirković, J. (2009). Incrementality and prediciton in human sentence processing. *Cogn. Sci.* 33, 583–609.

Arai, M., van Gompel, R., and Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cogn. Psychol.* 54, 218–250.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Pipenbrock, R., and Gulikers, L. (1995). *The Celex Lexical Database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The waCky wide web: a collection of very large linguistically processed web-crawled corpora. *Lang. Resour. Eval.* 43, 209–226.

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *J. Mem. Lang.* 59, 457–474.

Berkum, J. V., Brown, C. M., Zwitserlood, P., Koojiman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467.

Berry, M. J., Brivanlou, I. H., Jordan, T. A., and Meister, M. (1999). Anticipation of moving stimuli by the retina. *Nature* 398, 334–338.

Carminati, M. N., Gompel, R. P. G. van, Scheepers, C., and Arai, M. (2008). Syntactic priming in comrpehension: the role of argument order and animacy. *J. Exp. Psychol. Lang. Mem. Cogn.* 34, 1098–1110.

Chambers, C. G., Tanenhaus, M. K., and Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 687–696.

Chapman, C. S., Gallivan, J. P., Wood, D. K., and Milne, J. L. (2010). Reaching for the unknown: Multiple target encoding and real-time decision-making in a rapid reach task. *Cognition* 116, 168–176.

Chun, M. M., and Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychol. Sci.* 10, 360–365.

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye-movements. *Cogn. Psychol.* 42, 317–367.

DeLong, K., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121.

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211.

Federmeier, K. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* 44, 491–505.

Federmeier, K., McLennan, D. B., E. DeOchoa, E., and Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: an ERP study. *Psychophysiology* 39, 133–146.

Fischler, I., and Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *J. Verbal Learn. Verbal Behav.* 18, 1–20.

Hale, J. (2003). The information conveyed by words in sentences. *J. Psycholinguist. Res.* 32, 101–122.

Haskell, T. R., Thornton, R., and MacDonald, M. C. (2010). Experience and grammatical agreement: statistical learning shapes number agreement production. *Cognition* 114, 151–164.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., and Kliegl, R. (2011). dlexdb – eine lexikalische datenbank für die psychologische und linguistische forschung. *Psychol. Rundsch.* 62, 10–20.

Huettig, F., Olivers, N., and Hartsuiker, R. J. (2011a). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychol. (Amst.)* 137, 138–150.

Huettig, F., Singh, N., Singh, S., and Mishra, R. K. (2011b). "Language-mediated prediction is related to reading ability and formal literacy," in *Proceedings of the 17th Annual Conference on Architectures and Mechanisms for Language Processing*, Paris.

Kaiser, E., and Trueswell, J. C. (2005). The role of discourse context in the processing of a flexible word-order language. *Cognition* 94, 113–147.

Kamide, Y., Altmann, G. T. M., and Haywood, S. (2003a). The time course of prediction in incremental sentence processing. *J. Mem. Lang.* 49, 133–156.

Kamide, Y., Scheepers, C., and Altmann, G. T. M. (2003b). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *J. Psycholinguist. Res.* 32, 37–55.

Kaschak, M. P., Loney, R. A., and Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition* 99, B73–B82.

Knoeferle, P., and Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye-movements. *J. Mem. Lang.* 75, 519–543.

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition* 95, 95–127.

Koehn, P. (2005). "Europarl: a parallel corpus for statistical machine translation," in *Proceedings of MT Summit X*, Phuket.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). "The German reference corpus DeReKo: a primordial sample for linguistic research," in *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*, eds N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (Valletta: European Language Resources Association), 1848–1854.

Levy, R. (2008). Expectations-based syntactic comprehension. *Cognition* 106, 1126–1177.

Miall, R. C., Weir, D. J., Wolpert, D. M., and Stein, J. F. (1993). Is the cerebellum a smith predictor? *J. Mot. Behav.* 25, 203–216.

Nijhawan, R. (1994). Motion extrapolation in catching. *Nature* 370, 256–257.

Olivers, C. N., Humphreys, G. W., and Braithwaite, J. J. (2006). The preview search task: evidence for visual marking. *Vis. cogn.* 14, 716–735.

Pickering, M., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci. (Regul. Ed.)* 11, 105–110.

Rayner, K., and Raney, G. E. (1996). Eye-movement control in reading and visual search: effects of word frequency. *Psychon. Bull. Rev.* 3, 245–248.

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Saffran, J. (2003). Statistical language learning: mechanisms and constraints. *Curr. Dir. Psychol. Sci.* 12, 110–114.

Saffran, J., Aslin, R. N., and Newport, E. (1996). Statistical learning by 8-month old infants. *Science* 274, 1926–1928.

Scheepers, C. (2003). Syntactic priming of relative clause attachments: persistence of structural configuration in sentence production. *Cognition* 89, 179–205.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition* 71, 109–148.

Sloetjes, H., and Wittenburg, P. (2008). "Annotation by category – ELAN and ISO DCR," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Spivey, M. J., and Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychol. Res.* 65, 235–241.

Tulving, E., and Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *J. Exp. Psychol.* 66, 319–327.

Weber, A., Grice, M., and Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: a study of anticipatory eye movements. *Cognition* 99, B63–B72.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. C., and MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cognition* 58, 250–271.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269, 1880–1882.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: a mind/brain perspective. *Psychol. Bull.* 133, 273–293.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 August 2011; accepted: 28 November 2011; published online: 23 December 2011.

Citation: Knoeferle P, Carminati MN, Abashidze D and Essig K (2011) Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension. Front. Psychology ***2****:376. doi: 10.3389/fpsyg.2011.00376*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

APPENDIX

SENTENCE STIMULI

1. Der Versuchsleiter zuckert sogleich/zuckerte soeben die Erdbeeren/Pfannkuchen.
2. Der Versuchsleiter mixt sogleich/mixte soeben den Cocktail/Milchshake.
3. Der Versuchsleiter buttert sogleich/butterte soeben die Brotscheiben/Croissants.
4. Der Versuchsleiter bewässert nachher/bewässerte unlängst die Kresse/Tulpe.
5. Der Versuchsleiter poliert nachher/polierte unlängst die Kerzenständer.
6. Der Versuchsleiter studiert nachher/studierte unlängst den Buchtitel.
7. Der Versuchsleiter öffnet demnächst/öffnete kürzlich die Saftflasche/Schuhkiste.
8. Der Versuchsleiter würzt demnächst/würzte kürzlich die Gurke/Tomate.
9. Der Versuchsleiter salzt demnächst/salzte kürzlich die Zucchini/Aubergine.
10. Der Versuchsleiter schlürft baldigst/schlürfte vorhin die Limonade/Apfelschorle.
11. Der Versuchsleiter schüttelt baldigst/schüttelte vorhin die Sojamilch/Sprühsahne.
12. Der Versuchsleiter verrührt baldigst/verrührte vorhin den Milchkaffee/Kräutertee.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 11 June 2012
doi: 10.3389/fpsyg.2012.00156

Taking action: a cross-modal investigation of discourse-level representations

Elsi Kaiser*

Department of Linguistics, University of Southern California, Los Angeles, CA, USA

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Hannah Rohde, University of Edinburgh, UK
Ted J. M. Sanders, Universiteit Utrecht, Netherlands

****Correspondence:***
Elsi Kaiser, Department of Linguistics, University of Southern California, 3601 Watt Way, GFS 301, Los Angeles, CA 90089-1693, USA.
e-mail: emkaiser@usc.edu

Segmenting stimuli into events and understanding the relations between those events is crucial for understanding the world. For example, on the linguistic level, successful language use requires the ability to recognize semantic coherence relations between events (e.g., causality, similarity). However, relatively little is known about the mental representation of discourse structure. We report two experiments that used a cross-modal priming paradigm to investigate how humans represent the relations between events. Participants repeated a motor action modeled by the experimenter (e.g., rolled a ball toward mini bowling pins to knock them over), and then completed an unrelated sentence-continuation task (e.g., provided a continuation for "*Peter scratched John*…."). In two experiments, we tested whether and how the coherence relations represented by the motor actions (e.g., causal events vs. non-causal events) influence participants' performance in the linguistic task. (A production study was also conducted to explore potential syntactic priming effects.) Our analyses focused on the coherence relations between the prompt sentences and participants' continuations, as well as the referential shifts in the continuations. As a whole, the results suggest that the mental representations activated by motor actions overlap with the mental representations used during linguistic discourse-level processing, but nevertheless contain fine-grained information about sub-types of causality (reaction vs. consequence). In addition, the findings point to parallels between shifting one's attention from one-event to another and shifting one's attention from one referent to another, and indicate that the event structure of causal sequences is conceptualized more like single events than like two distinct events. As a whole, the results point toward common representations activated by motor sequences and discourse-semantic relations, and further our understanding of the mental representation of discourse structure, an area that is still not yet well-understood.

Keywords: psycholinguistics, discourse, causality, priming, coherence relations

INTRODUCTION

Our ability to segment stimuli into events and to understand the relations between those events is a key aspect of human cognition, and crucial for understanding and interacting with the world (e.g., Zacks and Swallow, 2007). Within the domain of cognitive psychology, there exists a large body of work investigating what cues humans use to recognize relations such as causality (e.g., Michotte, 1946/1963; Kanizsa and Vicario, 1968; Schlottmann et al., 2006) and similarity (e.g., Gati and Tversky, 1984; Gentner and Markman, 1997; Simmons and Estes, 2008). Many of these studies have focused on visual stimuli, such as the collision events used by Michotte and colleagues. However, as humans we also process information about events in other modalities, including language. In the linguistic domain, successful comprehension relies on listeners being able to recognize and understand the different kinds of relations that can hold between clauses (e.g., Hobbs, 1979; Mann and Thompson, 1986; Sanders et al., 1992; Kehler, 2002; Asher and Lascarides, 2003). For example, if someone says to a listener that "*Tom yelled at Peter*" and then continues with "*Peter kicked Tom's car*," the listener's understanding of what the speaker is trying to convey will be very different depending on whether she construes Tom's yelling to be what resulted in Peter kicking Tom's car (a causal relation), or whether she thinks Tom yelled at Peter because Peter had kicked his car (an explanation relation). In other words, the listener's inferences about the coherence relation between these two clauses (and correspondingly, the events they describe) have a fundamental effect on how she understands the situation. As noted by Webber et al. (2003), "a text means more than the sum of its component sentences. One source of additional meaning are relations taken to hold between adjacent sentences" (see also Sanders et al., 1993, p. 545). Thus, for successful communication, comprehenders need to be able to figure out the intended coherence relations between clauses[1].

[1]The research reported in this paper explores not only the relations that people construct between clauses (and the events they describe) but also the relations between events that are presented in a non-linguistic way. For the sake of brevity, when talking about people's interpretation of linguistic input, this paper will often simply refer to the *coherence relations between events*, rather than "the coherence relations between the events that are described by the two clauses" or "the coherence relations between

However, existing work in the linguistic domain has not reached a consensus about (i) what coherence relations there are, or (ii) how they are represented (see e.g., Sanders et al., 1993; Webber et al., 2003 for discussion). Some researchers argue that all coherence relations can be derived from a small set of primitives (e.g., Sanders et al., 1992; Kehler, 2002) whereas others work with a large, relatively unconstrained set of relations (e.g., Mann and Thompson, 1988). Furthermore, researchers differ in how they represent coherence relations, e.g., as hierarchical structures or as logical rules, and in what role they attribute to explicit connectives such as "because" and "as a result."

This paper aims to further our understanding of coherence relations – in particular, which relations are "psychologically real" and how they might be represented – by exploring the interface between the linguistic and non-linguistic domains. The experiments reported here used a cross-modal priming paradigm where participants carried out sequences of motor actions involving small objects, and then completed a seemingly unrelated linguistic sentence-continuation task. For example, a participant might roll a ball toward toy bowling pins in order to knock them over (cause-effect sequence), and then be asked to provide a continuation for a sentence such as "Peter tickled John." Two experiments tested whether and how the coherence relations represented by the motor actions (e.g., causal events vs. events that do not involve causality) influence participants' performance in the linguistic task.

Existing work has found evidence for action-language congruity effects in a range of areas, including the semantics of space and motion (e.g., Glenberg and Kaschak, 2002; Zwaan and Taylor, 2006; Glenberg et al., 2008) as well as emotional valence and motion (e.g., Casasanto and Dijkstra, 2010). For example, Zwaan and Taylor found that the physical act of rotating a knob interacts with the comprehension of sentences involving manual rotation, such as "Liza opened the pickle jar." These findings also receive support from neurolinguistic investigations showing that the cortical areas activated during the processing of action verbs such as "kick" overlap with the areas that are activated when people physically perform the same action (e.g., Buccino et al., 2005; Pulvermüller et al., 2005; Tettamanti et al., 2005). However, the question of whether discourse-level aspects of language production may also involve domain-general representations is not yet well-understood. For some early evidence, see Kaiser (2009), summarized in the General Discussion.

Both of the experiments reported here make use of priming – i.e., the observation that prior exposure to a stimulus influences (often facilitates) subsequent processing of a similar stimulus. Prior work has shown that priming occurs in a range of linguistic domains, including syntax, semantics, and phonology. For example, in the domain of syntax, producing a particular syntactic structure boosts the likelihood of the speaker producing the same structure again (e.g., Bock, 1986; Pickering and Branigan, 1998). The two experiments reported here use priming to see if two processes – the observation and execution of motor actions on the one hand, and language production on the other hand – make use of the same (or overlapping) underlying representations. Priming provides us with a tool to identify and diagnose properties of the representations utilized during the observation and execution of actions and during language production, which can further our understanding of the abstract mental representations involved in the production and conceptualization of coherence relations.

The experiments reported here have two common goals. The first goal is to learn more about how people represent coherence relations in the linguistic domain. As mentioned above, this is an area that is not yet well-understood, and many central questions remain open. The second goal is to learn about the relation between the linguistic domain and the non-linguistic domain, especially in terms of how humans represent relations between events in these two domains.

Experiment 1 tested whether (i) performing a motor action involving a cause-effect relation can bias participants to produce causal relations in a subsequent, unrelated linguistic task, and whether (ii) our mental representations distinguish between different sub-types of causality. If carrying out causal motor actions influences participants' linguistic choices in the production task, this provides evidence that the representations activated by the motor actions and discourse-level coherence representations overlap with each other. Furthermore, by taking a closer look at different kinds of causal relations – in particular the relation between causal sequences where the second action is an intentional reaction vs. causal sequences where the second action is an involuntary consequence – we can start to gain insights into what kind of information is encoded in these representations, i.e., how fine-grained they are.

Experiment 2 continues to explore the relation between linguistic and non-linguistic domains. Whereas Experiment 1 focuses on the question of whether fine-grained information about the relations between events can be represented in a domain-general way, Experiment 2 looks at a high-level, general property of events, namely event boundaries. This study has two main aims: first, to test whether the presence/absence of event boundaries in motor actions influences how participants complete the linguistic sentence-continuation task. In particular, it tests whether performing *two distinct motor actions* results in participants producing more continuations with *two distinct subjects* (i.e., continuations which shift attention to a new character), compared to a situation where only one motor action is performed. In other words, does shifting from one action to the next in one domain boost the likelihood of shifting from one referent to the next in another domain? We chose to analyze the subjects of participants' continuation sentences because of the well-known connection between subject hood and topicality (Reinhart, 1982; Chafe, 1994; Lambrecht, 1994). In other words, analyzing the subjects of the continuation sentences can provide a measure of topic-shifting, allowing us to assess whether shifting from one action to the next (in the domain of motor actions) has consequences on the linguistic level in terms of topic-shifts. Second, in order to gain insights into how causality is represented, Experiment 2 tests whether a causal action sequence patterns more like a sequence of two distinct actions or

the events that comprehenders assume the speaker intends their linguistic output to describe." However, despite this simplification, we assume that to fully understand linguistic input, comprehenders construct a propositional representation of the events that a particular linguistic form describes and also engage in real-world reasoning about the current state of affairs (e.g., that kicking someone's car might result in the car owner reacting negatively.)

like a single action. Because causal sequences often consist of multiple sub-events (e.g., event 1: I roll the ball, event 2: the bowling pins fall over), it is not *a priori* clear whether they are conceptualized as a single event (possibly with complex internal structure) or as two separate events.

Broadly speaking, the research presented in this paper has implications for our understanding of the mental representation of coherence relations, an area that is not yet well-understood. The results suggest that motor actions activate richly encoded representations that can overlap, on an abstract level, with discourse-level aspects of language. Investigating effects of motor actions on language further contributes to our understanding of causality sub-types and how causal sequences are conceptualized.

EXPERIMENT 1

Experiment 1 focuses on two related issues, namely (i) the domain-generality of coherence representations and (ii) the level of detail present in these representations. In exploring the domain-generality of how people represent relations between events, this study focuses on the notion of causality. Causal connections have been argued to be fundamental to how humans conceptualize events (e.g., Sanders, 2005; see also Trabasso and van den Broek, 1985; Wolfe et al., 2005 on the facilitative effects of causal connections on memory and processing), and Experiment 1 tests whether causal relations between physical events involve the same kinds of mental representations as causal relations between linguistically encoded events. Specifically, Experiment 1 tests whether execution of motor actions that represent causal relations influences the rate of causal relations produced in a language task.

In addition, this study also asks how detailed such causality representation are. Do comprehenders merely activate a rudimentary notion of causality that is shared across domains, or does this domain-general representation include fine-grained information about sub-types of causality? In particular, this study focuses on the distinction between two sub-types of causality: (i) situations where the result is involuntary consequence and (ii) continuations where the result consists of a volitional, intentional reaction. In the subsequent discussion, these two causal sub-types are referred to as the *consequence-type* and the *reaction-type*. Examples are shown in (1). In (1a), the result of falling over is an involuntary consequence of being kicked, whereas in (1b), the act of kicking back is a deliberate, intentional reaction to the original kicking event.

Jason kicked Matt. Matt fell over. ⇒ consequence type (1a)

Jason kicked Matt. Matt kicked him back. ⇒ reaction type (1b)

Although most linguistic approaches to coherence relations do not distinguish these two sub-types of causality, this distinction is made in Rhetorical Structure Theory (RST, Mann and Thompson, 1988), a theory which aims to provide a descriptive characterization of how text is organized. Mann and Thompson propose a large number of different discourse relations, including "Volitional result" and "Non-volitional result." The former is (i) a situation where the initial action/situation causes another action that is volitional, whereas the latter is (ii) a situation where the initial action/situation causes another action that is not volitional (see Mann and Thompson, 1988, p. 275 for further details and examples). Thus, this corresponds to the distinction between consequence-type and reaction-type causal relations. Recent research on Dutch by Stukker et al. (2008) also makes a number of important, fine-grained distinctions regarding sub-types of causality, including intentional vs. non-intentional causation (see e.g., Stukker et al., 2008, p. 1305 regarding the use of the two connectives *daardoor* "because of that" and *daroom* "that's why," which are associated with non-intentional and intentional causality, respectively).

However, as Knott (1993) notes, it is important to ask whether this distinction is psychologically real: "How do we decide whether to subdivide or not to subdivide result into volitional result and non-volitional result? Again, different cuts through the space of relations are possible: why distinguish between volitional and non-volitional result, and not between, say, immediate and delayed result?" (Knott, 1993, p. 48). Shedding light on this question is the second main aim of Experiment 1. Thus, in addition to investigating the domain-generality of causality representations, this experiment also tests whether the distinction into reaction-type causality and consequence-type causality is justifiable, and in doing so, aims to gain new insights into how detailed our representations of causality are.

MATERIALS AND METHODS

Participants

Thirty adult native English speakers from the University of Southern California community participated. All studies reported in this paper were approved by the University of Southern California University Park Institutional Review Board, which is fully accredited by the Association for the Accreditation of Human Research Protection Programs (AAHRPP).

Materials

Motor action trials (Priming trials). This study used 12 critical prime actions and 24 filler actions. The actions involved manipulating small toys or other objects. The critical actions were of three types: (i) Causal actions, (ii) Two-Event actions, and (iii) One-Event actions. In causal actions, one action causes something to happen (e.g., rolling a ball into dominos to make them fall over). Because the prime actions all involved inanimate objects/toys, the Causal actions all exemplify consequence-type causality. Two-event actions involved two distinct actions that are not causally connected (e.g., open and close a folding ruler, tie a bendy pencil into a knot). One-event actions involved events that could be construed as a single action (e.g., building part of a jigsaw puzzle). Examples are provided in **Table 1** and **Figure 1**.

Norming study. An initial norming study was conducted to ensure that the three action types were indeed conceptualized as intended. The norming study included a large set of different actions, including Causal actions, Two-Event, and One-Event actions, actions that involved sorting objects into categories, and other kinds of actions. Eighteen native English speakers (who did not participate in any of the other studies) watched the experimenter perform each action, repeated the action themselves and were then asked to indicate whether the action is best described as "two unrelated things happening," "one thing causing

another thing to happen," "two similar things happening," "objects being sorted into different categories or groups," or "none of the above." Based on the outcomes of this norming study, four action sequences that were consistently judged to be causal were identified and chosen as the Causal primes for the main experiment. Furthermore, four action sequences that were consistently judged to involve two unrelated things happening were chosen to be the Two-Event actions in the main experiment, and four One-Event action sequences were chosen from actions that in the priming study were not judged to involve causation, similarity, sorting, or multiple actions.

Table 1 | Examples of prime actions.

Causal	Roll a ball toward domino pieces to knock them over (**Figure 1A**)
Causal	Push a toy car so that it runs into a second toy car and makes the second car move forward
One-event	Assemble a corner of a jigsaw puzzle (**Figure 1B**)
One-event	Build a sandwich using toy/fake "food"
Two-event	Open and close folding ruler, tie a knot in bendy pencil (**Figure 1C**)
Two-event	Make an X-shape with two yellow sticks, then roll a die

It is worth noting that the One-Event actions involve smaller sub-actions (e.g., combining the different jigsaw pieces into a bigger piece of the puzzle). Thus, the term "One-Event" refers to the cumulative event that is composed of the smaller sub-actions. These kinds of One-Event actions were used in order to keep the duration and intuitive "complexity" of the actions as comparable as possible. Crucially, people's norming responses suggest that they did not perceive the One-Event actions as involving causality, similarity or two distinct events.

Generally speaking, any action or event can be viewed on different levels of granularity and decomposed into smaller and smaller sub-parts (e.g., the act of picking up a puzzle piece could be further decomposed into various sub-components involving visual perception, programming of a reaching motion, carrying out the reaching motion, and so on). What is most relevant here is that, relatively speaking, the sub-components of the One-Event trials are conceptualized as contributing toward a single goal-driven action (e.g., assembling a puzzle or building a sandwich). Thus, in this regard they contrast with the Two-Event actions, which do not form a single, coherent, goal-driven action.

The 24 filler actions used in the main study were also chosen on the basis of the norming study, to ensure that they were not perceived as involving causation, similarity, sorting or multiple

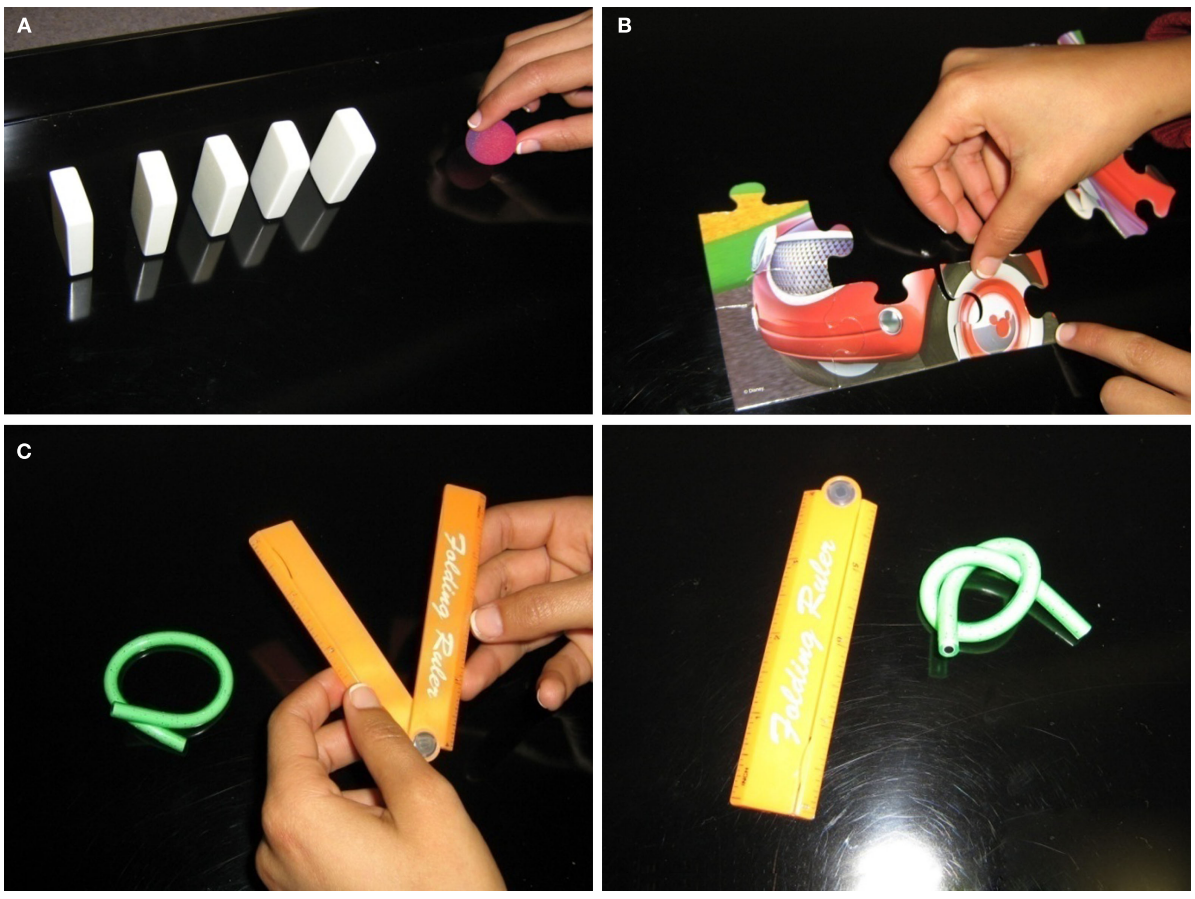


FIGURE 1 | (A) Example of a Causal action. **(B)** Example of a One-Event action. **(C)** Example of a Two-Event action.

actions. This was done to minimize any danger of the filler actions priming the target trials.

Sentence-continuation trials (Target trials). In the main experiment, the motor action prime trials were intermixed with semantically unrelated sentence-continuation trials, where participants provided a continuation sentence to a transitive prompt sentence (ex.2). The study contained 9 critical sentence-continuation trials and 24 filler sentence-continuation trials. The critical prompt sentences were transitive sentences with two male or two female names. The verbs were all agent-patient verbs involving physical interaction (*kick*, *pinch*, *tickle*, *scratch*, *slap*, *punch*, *poke*, *push*, *hit*)[2]. Agent-patient verbs were used in order to keep the semantic class of the verbs consistent. We chose to use verbs involving physical interactions because the prime actions were also physical (see e.g., Schlottmann et al., 2006; on differences between physical and non-physical causation, see also Kanizsa and Vicario, 1968).

Critical trials consisted of pairs of action primes and sentence-continuation prompts. Because there were three conditions (Causal prime action, Two-Event prime action, and One-Event prime action), we created three lists using a Latin-Square design. Reverse versions of each list were also created to control for effects of presentation order. In both forward and reverse lists, the prime action trial immediately preceded the sentence-continuation trial (i.e., there were no intervening trials between primes and targets). However, the filler actions and filler sentence-continuation trials were not presented in pairs, but rather pseudorandomly intermixed. This was done to ensure that participants would not perceive the actions and the sentences as being connected to each other.

Jason kicked Matt. `Matt hit him in retaliation.`
Jason kicked Matt. `He was a rather violent person.` (2)

PROCEDURE

Participants sat in front of a computer screen at a wide table. On action trials, the screen showed the word "ACTION." Upon seeing this, the participant turned away from the computer screen, watched the experimenter perform the intended action, and then repeated it. (The actions were not described in words at any point.) After completing the action, the participant would press a key on the keyboard. The screen would then move on to the next trial, and show the word ACTION (if the next trial was also an action trial), or it would show a sentence that the participant had to type a continuation for (if the next trial was a sentence-completion trial). On sentence-continuation trials, participants were instructed to write a natural-sounding continuation sentence for the sentence shown on the screen. They were encouraged to avoid overthinking, and to write what first came to mind. After completing the sentence-continuation, the participant would press a key, and the screen would either show the word ACTION or another sentence that the participant was asked to continue. This set-up was used to create the impression that the ordering of sentence-completion trials and action trials was random.

Coding

Continuations were double-coded by two blind coders, who analyzed the semantic coherence relation between the prompt sentence and the continuation sentence provided by the participant. For example, coders marked whether the event in the continuation sentence was a consequence of the event described in the prompt sentence, or perhaps an explanation why the prompt sentence event happened. The coding schema used the coherence relations from Kehler (2002) and Kehler et al. (2008). The relations that are most relevant to the current discussion are shown in **Table 2**, with examples. Building on Rohde (2008), training and detailed coding guidelines were used to ensure consistency among coders. Each coder went through the data independently. Coders were instructed to be conservative and to avoid over-interpretation, i.e., to err on the side of choosing "unclear" if there was not enough information available to determine the intended coherence relation. Subsequently, any discrepancies between the coders were resolved through discussion. If the two coders did not agree on a coherence relation or agreed that not enough information was available to determine intended coherence relation, the trial was coded as "unclear." In the end, 4.8% of the critical trials were coded as having unclear/ambiguous relations.

In addition to the coherence relations from Kehler's work, we also distinguished two sub-types of cause-effect relations, as mentioned above: (i) continuations where the result is involuntary/automatic consequence (consequence-type) and (ii) continuations where the result consists of a volitional, intentional reaction (reaction-type). There were also some continuations that were judged to be causal but it was unclear which of these two groups they belonged to. These were coded as a third sub-type, "unclear causal" (e.g., *Joe punched Tom. Tom resented Joe for the result of his life.* Here, Tom's resenting Joe is caused by the punch, but it is not clear whether should be regarded as an involuntary, automatic response or – especially in light of the long duration of the resentment – as a more volitional reaction.) In the end, 5.6% of causal continuations were coded as "unclear causal."

Predictions

This experiments tests two main predictions. The more general prediction has to do whether causal actions will boost the rate of causal continuations in the sentence-completion task. If causal action primes result in more causal continuations, this indicates that these two processes make use of the same (or overlapping) underlying representations.

The second main prediction has to do with the level of detail that is encoded in the relevant representations. Importantly, the causal motor action primes used in Experiment 1 only involve involuntary consequences (e.g., *the bowling pins fall*

[2]The study also included three verbs of social interaction (*embrace*, *greet*, and *hug*). However, these verbs were excluded from further analysis because their semantic properties differ from the agent-patient verbs. In particular, these social interaction verbs are (by default) construed as involving reciprocal actions – for example, when someone greets another person, the default assumption is that the other person reciprocates. Crucially, these verbs do not involve a clear agent-patient asymmetry, in contrast to verbs like "hit" or "scratch" where one person is clearly the agent and other is the "undergoer" who is affected by the agent's actions (see also Levin, 1993 for more on verb classes).

Table 2 | Some of the most important coherence relation labels used in coding, and examples from participants' continuations.

Causal sub-type – consequence: the event in the second sentence was caused by the event described in the first sentence, but the consequence is involuntary, automatic	(i) Jason kicked Matt. `Matt felt hurt.` (ii) Lisa pinched Nancy. `Nancy immediately woke up.`
Causal sub-type – reaction: the event in the second sentence was caused by the event described in the first sentence, and the second event is an intentional, voluntary (re)action to the first event	(i) Greg slapped Josh. `Josh punched him back.` (ii) Tony hit Kevin. `Kevin called the police.`
Explanation: the second sentence provides an explanation of why the event in the first sentence happened ("because")	Angela scratched Melissa. `Melissa's back was itching.`
Elaboration: The second sentence provides a restatement of the first sentence, perhaps from another perspective or with more information	Ken poked Steven. `He poked Steven right in the gut.`
Occasion: the second sentence describes an event that happens after the event described in the first sentence, but is not caused by the event in the first sentence. ("narrative" relation)	William tickled David. `William took a video of David's laughing fit and put it on YouTube.`

over). There are no causal primes with results that were volitional reactions. Thus, by looking at which sub-type participants' causal continuations fall into, we can see whether the motor primes' consequence-type nature is mirrored in the linguistic continuations. If yes, this suggests that the representations that overlap are more detailed than a simple causal/non-causal division might suggest, i.e., that the domain-general representation of causality is nevertheless sophisticated enough to include the distinction between consequence-type and reaction-type causal relations.

Both One-Event and Two-Event prime actions were included in order to check whether the simple number of events could play a role. In particular, it could be that what is being primed is the number of distinct events or predicates or the fact that there is a temporal sequence such that the second event occurs after the first event. According to this view, if someone carried out a two-event action – regardless of whether it's causal or non-causal – this might prime them to produce a causal continuation rather than an explanation or an elaboration, for example. Thus, in order to be able to probe whether causal actions in particular are priming causal continuations in the sentence-completion task, Experiment 1 included both One-Event actions and non-causally related Two-Event actions (in addition to the Causal actions).

RESULTS

To test whether the prime actions influenced participants' continuations, mixed-effects logistic regression models were used (e.g., Baayen et al., 2008) to analyze (i) the o*verall proportion of causal continuations* as a function of condition (Causal, One-Event, Two-Event), (ii) the proportion of *consequence-type causal continuations* as a function of condition, and (iii) the proportion of *reaction-type causal continuations* as a function of condition. In the initial general analyses, the three sub-types of causal continuations – reaction-type causal continuations, consequence-type causal continuations and unclear causal continuations – were all grouped together. In each analysis, participant, and item were included as random effects[3]. Mixed-effects regression models were used because the data is categorical and thus not well-suited for ANOVAs (see e.g., Jaeger, 2008). At the end of the results section, we also consider a production study that addresses the question of whether the results of Experiment 1 could be attributed to syntactic priming. As will become clear, we argue that this is not the case.

GENERAL CAUSALITY

Starting with the overall proportion of causal vs. non-causal continuations, we see in **Figure 2A** (which includes all three sub-types of causal continuations; consequence causal, reaction causal and unclear causal) that participants' continuations do indeed show significant effects of prime type: after performing Causal actions, participants produced significantly more Causal continuations than after performing One-Event actions (80 vs. 66%, $\beta = -1.115$, Wald $Z = -2.599$, $p < 0.01$). The rate of Causal continuations after Causal actions was also marginally higher than the rate of Causal continuations after Two-Event actions (80 vs. 71%, $\beta = -0.774$, Wald $Z = -1.811$, $p = 0.07$). In contrast, there is no significant difference between the rate of Cause continuations occurring after One-Event and Two-Event actions ($\beta = 0.227$, Wald $Z = 0.621$, $p = 0.53$). In sum, the results indicate that performing a Causal action makes participants more likely to produce a causally connected continuation in the sentence-continuation task, as compared to non-Causal actions.

CAUSALITY SUB-TYPES

When we take a more detailed look at the two kinds of causal sub-types, consequence-type causality and reaction-type causality, a striking asymmetry emerges in terms of whether their frequencies are affected by the action primes. First, when one considers the proportion of reaction-type causal continuations, shown in **Figure 2B**, there are no effects of priming (ps > 0.5). In general, the rate of reaction-type continuations is relatively low (below 30%).

[3]When specifying the structure of random effects, we started with fully crossed and fully specified random effects, tested whether the model converges, and reduced random effects (starting with item effects) until the model converged (see Jaeger at http://hlplab.wordpress.com, May 14, 2009). Then, we used model comparison to test each random effect; only those that were found to contribute significantly to the model were included in the final analyses. However, all models contained random intercepts for subjects and items.

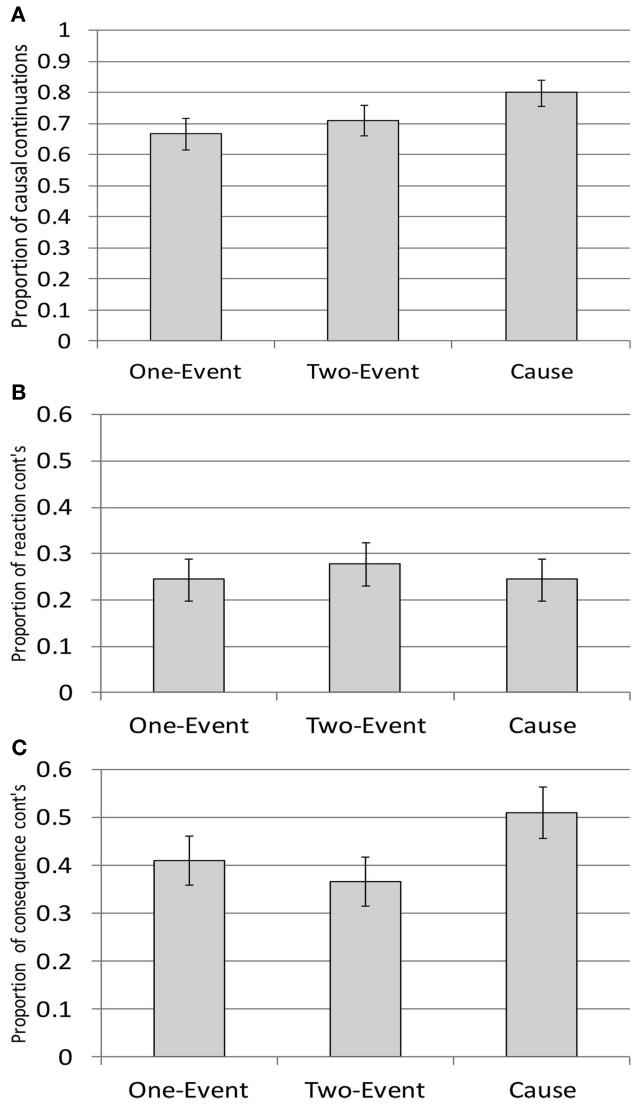


FIGURE 2 | (A) Proportion of causal continuations as a function of the three different kinds of prime actions. (Error bars show ± 1 SE). **(B)** Proportion of *reaction-type continuations* as a function of the three different kinds of prime actions. (Error bars show ± 1 SE). **(C)** Proportion of *consequence-type continuations* as a function of the three different kinds of prime actions. (Error bars show ± 1 SE.).

In contrast, the rate of consequence-type causal continuations – as shown in **Figure 2C** – is clearly affected by the kind of motor action that participants performed during the priming trials: there are significantly more consequence-type continuations after Causal actions than after Two-Event actions (51 vs. 36.7% consequence continuations, $\beta = -0.83$, Wald $Z = -2.466$, $p < 0.02$). Similarly, the rate of consequence continuations was numerically higher after Causal actions than after One-Event actions (51 vs. 41%, $\beta = -0.486$, Wald $Z = -1.542$, $p = 0.123$). The rate of consequence-type continuations after One-Event actions and Two-Event actions did not differ significantly ($\beta = -0.269$, Wald $Z = -0.84$, $p = 0.4$).

COULD THESE EFFECTS BE DUE TO SYNTACTIC PRIMING? PRODUCTION STUDY

A potential question that comes up is whether the effects observed here could be due to syntactic priming. It is well-known that hearing or producing a particular kind of syntactic structure makes people more likely to produce that structure again (Bock, 1986; Pickering and Branigan, 1998; Bock and Griffin, 2000; Arai et al., 2007). In Experiment 1, the primes were presented in a non-linguistic modality, but participants were not prevented from encoding them linguistically (e.g., silently describing the action in words). Thus, one might wonder whether the results reported in the preceding sections could be due to priming of syntactic representations.

To address this question, a production study was conducted: 24 new participants watched the experimenter carry out the action, repeated the action themselves, and were then asked to describe the action in words. More specifically, participants received the following instructions: "What did you do/what happened? You should write down whatever you feel best describes what happened, using whatever words seem most appropriate to you." Afterward, the syntactic properties of participants' descriptions were analyzed. As will become clear below, the results of the production study show that it is very unlikely that the results described above are due to syntactic priming.

For *One-Event actions*, 92.7% of people's descriptions were monoclausal structures (e.g., "I arranged the sticks in a hexagon," "I arranged puzzle pieces for the bottom left corner of a puzzle"). The rest were also one-clause descriptions but contained an additional fronted clause (e.g., "Using 6 puzzle pieces, I completed a portion of a puzzle of a red car"). In contrast, participants' descriptions of *Two-Event actions* always included two verbs/two predicates, due to the semantics of these primes involving two distinct actions (e.g., "I made a cross out of two yellow sticks and then rolled a red die," "I put two sticks on top of each other and then rolled a die," "I opened and closed an orange folding ruler before tying a knot into a piece of green sparkly plastic tube.")

However, the descriptions of the *Causal actions* are the ones that are most relevant for the question of whether the results of Experiment 1 could be due to syntactic priming. For the Causal actions, 97.9% of descriptions were highly transitive (i.e., include a subject, a verb, and a direct object), for example "I positioned five dominoes in a line and knocked them over with a rubber ball," "I stacked the dominoes in a row, and knocked them down with the ball," "I placed two green and red toy cars facing right and pushed the red one to hit the green and move it[4]." All but three of these descriptions included two or more transitive clauses

[4]Could the highly transitive nature of the causal descriptions be an artifact caused by the instructions and not a true reflection of how the participants in Experiment 1 might have verbalized the events? Specifically, could it be that the production instructions caused an artificially high rate of responses like "I knocked over the dominoes with a ball" instead of "The ball knocked over the dominoes"? In our opinion, the high rate of first-person sentences is unlikely to be due to the wording of the instructions: First, in both Experiment 1 and the production study, each trial involved two occurrences of the action being conducted by a human agent (the experimenter and then the participant), and so it seems unlikely that participants would verbalize the actions without encoding the human agent. The second

(like the examples above), and the three remaining descriptions consisted of a single transitive clause (e.g., "I pushed the red car into the green car.") In sum, in the vast majority of cases people produced multiple transitive clauses when describing the causal actions.

The high rate of transitive sentences is noteworthy when coupled with the observation that consequence-type continuations are *less transitive* than reaction-type continuations. Consequence-type continuations (where the result is an involuntary consequence) are often intransitive and lack a direct object (e.g., "She felt hurt", "He fell over"), whereas reaction-type continuations (where the result is a volitional, intentional reaction) are often highly transitive and mention an object to whom some action is done (e.g., "He punched him back", "Melissa told on Angela"). If the causal actions were syntactically priming participants' continuations in the sentence-completion task, Experiment 1 should have resulted in the exactly *opposite* pattern of what was actually found, namely Causal actions boosting the rate of consequence-type continuations but having no effect on the rate of reaction-type continuations.

In sum, we take the results of this production study as an indication that the results of Experiment 1 cannot be attributed to syntactic priming: if anything is being primed by the causal actions, it is a transitive structure, which could not generate the results that were obtained. (It is important to note that the aim of the production study was simply to address potential concerns regarding syntactic priming. The finding that the actions used in Experiment 1 were almost always described with transitive sentences should not be interpreted as a claim that *all* causal event sequences must be described with transitives; the relation between transitivity and event structure is a complex topic that is beyond the scope of this paper. The modest aim of the production study was simply to assess the potential impact (or lack thereof) of syntactic priming.) In sum, it seems reasonable to conclude that the results of Experiment 1 cannot be attributed to priming on the level of syntactic representations. Rather, it seems that priming is taking place on the level of more abstract conceptual representations that are shared both by motor actions and linguistic representations.

DISCUSSION

Experiment 1, which used a priming paradigm involving motor actions that preceded target trials in a sentence-completion task, showed that Causal action primes resulted in more causally connected sentence-completions than One-Event or Two-Event action primes. As a whole, this finding points toward a shared abstract level of representation being activated/used by motor sequences and discourse-level coherence relations.

More specifically, the results show that the priming effect is carried by an increase in the rate of consequence-type causal continuations, and not the rate of reaction-type continuations: participants were equally likely to produce reaction-type continuations in all three prime conditions. In contrast, after carrying out a causal action sequence involving a consequence-type relation, participants produced a higher rate of consequence-type continuations in the sentence-completion task, compared to non-causal action primes. Overall, causal primes resulted in significantly more consequence-type continuations than Two-Event primes and in numerically more consequence-type continuations than One-Event primes. As predicted, One-Event primes and Two-Event primes do not differ in the proportion of subsequent consequence-type continuations. (It is not clear why the difference between Causal primes and One-Event primes does not quite reach significance.)

Given that the causal motor actions involved consequence-type relations rather than reaction-type relations, the results of Experiment 1 suggest that a shared abstract level of representation is activated by motor sequences and discourse-level coherence relations, and that this level of representation is sufficiently detailed to encode the distinction between consequence and reaction. However, it is important to keep in mind that participants were not prevented from encoding the prime actions linguistically (i.e., were not prevented from putting them in words). Thus, the motor action information could have been converted into some kind of linguistic information by the participants, which in turn could be what overlaps with the representations that participants use in the sentence-continuation task. Importantly, the production study described above provides evidence that the results of Experiment 1 cannot be derived from syntactic priming. This indicates that the relevant level of representation is not syntactic. Instead, it seems more plausible to assume that the motor actions, whether they are encoded linguistically or not, are activating semantic representations that also involve information about the relations between events, and that this is what overlaps with the representations used in the sentence-continuation task.

In sum, the results of Experiment 1 indicate that causality representations, even when *originating from non-linguistic, motor action input*, seem to be sufficiently richly encoded to have subtle effects on language production.

It is interesting to note that the distinction between reaction and consequence is also relevant in the domain of cognitive psychology for the difference between physical causation and social causation (e.g., Kanizsa and Vicario, 1968; Schlottmann et al., 2006). A situation where one billiard ball hits another involves *physical causation*, whereas a situation where one animal runs away from another involves *social causation*. Although not normally described in terms of reaction vs. consequence, it seems that the distinction between physical and social causation could be interpreted as mapping onto the distinction between consequence-type causal relations and reaction-type causal relations respectively. Work in cognitive psychology suggests that there are some differences in the perception of social and physical causality by adults (Schlottmann

(related) reason why we expect people's descriptions to have human agents regardless of the instructions is based on prior work showing that animate entities (in this case "I," the participant) are highly accessible and usually realized in subject position (e.g., Branigan et al., 2008). Thus, it seems that the high rate of transitive sentences with first-person subjects is unlikely to be due to the instructions. We feel that the production study can be used to test whether syntactic priming might be responsible for the results of Experiment 1 (and to argue, as we do, that the results cannot be attributed to syntactic priming). Further evidence for the claim that the results discussed in this paper cannot be reduced to syntactic priming comes from Experiment 2, where participants produced an increased proportion of sentences with two *distinct* subjects following Two-Event primes – which, if described linguistically, would yield two sentences with the *same* subject (e.g., "**I** opened the ruler and **I** tied the pencil in a knot.").

et al., 2006), and this seems to align well with the results of Experiment 1, which point to a cognitively meaningful distinction between reaction-type causal relations and consequence-type causal relations.

EXPERIMENT 2

The results of Experiment 1 indicate that detailed information about coherence relations – causality in particular – can be represented in a domain-general way. This suggests that our mental representations of coherence relations contain fine-grained, specific information. However, if our aim is to learn more about the mental representations of coherence relations, we also want to gain an understanding of the more general properties of coherence representations. Thus, Experiment 2 shifts away from the specifics to a more abstract level, and explores a very general property of event sequences, namely the representation of event boundaries. The human ability to segment stimuli into distinct events is a crucial aspect of cognition. In the visual domain, the boundaries between events have been shown to have effects on attention and memory (e.g., Swallow et al., 2009), suggesting that the cognitive process of shifting from one event to another has far-reaching effects on humans' mental representations.

Experiment 2 addresses two main questions. First, it tests whether the presence/absence of event boundaries in the domain of motor actions influences how participants complete the linguistic sentence-continuation task. In particular, does performing two distinct actions (Two-Event primes) make participants more likely to produce continuations with two distinct subjects – i.e., continuations which shift attention to a new character? Conversely, does performing one action (One-Event primes) make participants more likely to maintain focus on the subject of the prompt sentence? As will be discussed in more detail in the "predictions" section, referent shifts were used as the dependent variable because of the well-known association between subjects and topics. The second main aim of this experiment is to gain insights into how causality is represented, and so it tests whether Causal primes pattern like Two-Event or like One-Event primes, in terms of the referential shift patterns that they induce. In other words, how are two causally connected events conceptualized – more like a one-event situation or a two-event sequence? A better understanding of this issue can help to clarify whether the event structure of causal sequences is best grouped with one-event representations or two-event representations.

MATERIAL AND METHODS

Participants

Twenty-four adult native English speakers from the University of Southern California community participated. None of the participants had participated in the other studies reported in this paper.

Materials

Motor action trials (prime trials). The same actions were used as in Experiment 1.

Sentence-completion trials (target trials). Instead of the agent-patient verbs used in Experiment 1, this experiment uses a class of so-called implicit-causality verbs (IC; Garvey and Caramazza, 1974; Stewart et al., 2000; Koornneef and Van Berkum, 2006), namely so-called Noun1 IC verbs, e.g., *frighten*, *annoy*, and *amuse*:

Angela frightened Melissa.

`She was wearing a scary mask.` (3)

Noun1 IC verbs were chosen for Experiment 2 because they allow for a situation where *no overwhelming subject or object bias* is expected, and thus they are well-suited for the purpose of testing whether the prime motor actions can induce referential shifts. The agent-patient verbs used in Experiment 1 would not have been suitable for this purpose, because they have a strong preference to shift to talking about the object (e.g., Stevenson et al., 1994), which could mask weak shifts toward the subject or lead to potential ceiling effects in the case of the object.

Let us consider in more depth why Noun1 IC verbs are well-suited for Experiment 2: prior work has shown that, when followed by the connective *because*, this particular class of IC verb tends to elicit continuations that start with reference to the preceding subject. For example, with a sentence like "Angela frightened Melissa because ..." or "Angela amused Melissa because ...," the presence of the connective "because" signals that an explanation must be provided, and so participants tend to continue by saying something about the subject Angela (Noun1). Given this robust preference, this class of IC verbs is called Noun1 IC verbs. However, Rohde (2008, see also Kehler et al., 2008; for summary of these results) showed that when *no overt connective* is provided and participants' continuations constitute a new sentence (as in ex.3), Noun1 IC verbs show a very different pattern: now, continuations after Noun1 IC verbs are almost equally likely to refer to the preceding subject or the preceding object (about 60% subject continuations, 40% object continuations) – this differs strikingly from the pattern that is observed with a "because" connective (85% subject continuations, see Rohde, 2008). The absence of a clear subject preference in the absence of an explicit connective presumably stems from the resulting absence of any explicit coherence relation constraints: when given a sentence with an IC verb that is not followed by an explicit *because*, participants still produce a fairly high rate of explanation continuations (over 55% in Rohde's study), but they also produce other coherence relations, many of which tend to start with the non-subject. This shifts the overall reference pattern to one where the preceding subject and object are (near)equal candidates for subsequent reference. Thus, thanks to this balanced situation, Noun1 IC verbs with no overt connective are well-suited for Experiment 2, where we are interested in seeing whether priming with motor actions influences the likelihood of maintaining vs. shifting reference.

It is worth noting that although explanation relations resemble causal (cause-effect) relations in that both refer to causes and consequences – albeit in a different linear order –, existing research suggests that these relations differ in fundamental ways. Causal relations are often regarded as more iconic, since they reflect the natural chronological order of events, unlike explanation relations (see van den Broek, 1990; see also Zwaan and Radvansky, 1998).

This fundamental distinction is supported by recent psycholinguistic research by Briner et al. (2012) who found that explanation relations are processed more slowly than causal relations (see also Noordman, 2001 for related work and Johnston and Welsh, 2000 for data from language acquisition). In light of these differences, this paper treats causal relations and explanation relations as distinct.

Coding. Continuations were analyzed for which character is mentioned at the start of the continuation, the preceding subject or object (or both or neither). As in Experiment 1, two coders blind to the experimental conditions worked independently. Afterward, disagreements were resolved through discussion. If the coders could not agree, the item was marked as "unclear" (13.5% of the trials).

Predictions

When continuing the prompt sentence, e.g., "Jason frightened Matt", participants may opt to continue by talking about Jason, as shown in ex(4a). Here, the prompt sentence and the continuation sentence have the same subject, Jason. In other words, we are maintaining focus on the initial subject. This referent maintenance pattern can also be thought of as topic maintenance, given that the topic of a sentence is normally realized in subject position in English (Reinhart, 1982; Chafe, 1994; Lambrecht, 1994).

Alternatively, participants may choose to shift to talking about the other character, namely the preceding object [ex (4b)]. Here, the prompt sentence and the continuation sentence have different subjects, in a pattern that we can characterize as shifting to a new character or topic-shift.

Jason frightened Matt. **He** was wearing a scary mask. (4a)

Jason frightened **Matt. He** ran away screaming. (4b)

Experiment 2 aims to test whether the presence/absence of event boundaries in motor actions influences the likelihood of topic-shifts in the sentence-completion task. If the mental representations activated by shifting from one motor action to another (Two-Event primes) overlap with the representations activated when shifting from one referent to another [topic-shift, ex.(4a)], then there should be more topic-shifts (object-referring continuations) after Two-Event primes than after One-Event primes. In other words, we should find a higher rate of object-referring continuations (and a lower rate of subject-referring continuations) after Two-Event primes than after One-Event primes.

If this holds, it would show that in a situation where the prime clearly involves two distinct events, this can induce referent shifts. With this finding, we can then turn to the Causal primes, to see whether they pattern more like Two-Event primes or like One-Event primes. In other words, will they trigger shifts to another character, or will we see a pattern of topic maintenance? The former result would suggest that Causal sequences are conceptualized as two events, whereas the latter would indicate that Causal sequences are conceptualized as single events.

More broadly, if Experiment 2 reveals an effect of the prime actions' event structure on the referent-shift/referent maintenance patterns in participants' continuations, this would provide evidence that the representations activated by the event structure of the motor actions overlap with the representations that are activated during referential processing in discourse.

RESULTS AND DISCUSSION

Mixed-effects logistic regression was used to test whether the rate of object-referring continuations differs as a function of the prime action. In other words, are people more likely to shift to talking about the object of the preceding sentence after some prime types than others? Only those continuations that started with reference to either the preceding subject or object were included; continuations that were coded as "unclear" or that began with reference to another entity were excluded from subsequent analyses (13.5%). In each analysis, participant, and item as were included as random effects[5].

As can be seen in **Figure 3**, the proportion of continuations that started with the prompt sentence object is significantly higher after Two-Event primes than after One-Event primes (65.9 vs. 45.24%, $\beta = 1.121$, Wald $Z = 3.181$, $p < 0.005$). This suggests that the rate of referential shifts does indeed correlate to the One-Event vs. Two-Event distinction in the predicted way: when the prime action shifts from one event to another, this is reflected in participants' continuations. Crucially, a comparison of Two-Event primes and Causal primes reveals significantly more shifting to the prompt sentence object after Two-Event primes than after Causal primes (65.9 vs. 48.75%, $\beta = -1.098$, Wald $Z = -2.796$, $p < 0.01$). Causal primes and One-Event primes did not differ ($\beta = 0.303$, Wald $Z = 0.844$, $p = 0.4$). In sum, Causal prime actions pattern like One-Event prime actions, and both of these differ from Two-Event primes.

We interpret the finding that Two-Event primes result in more object-referring continuations than One-Event primes as evidence

[5]Random effect structure was determined as in Experiment 1.

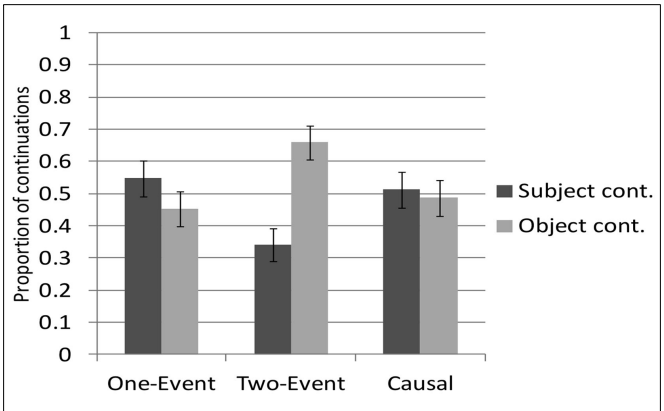


FIGURE 3 | Proportion of continuations that start by referring to the preceding subject or the preceding object, as a function of prime type. (Error bars show ± 1 SE.)

that the mental representations activated by shifting from one-event to another event overlap with the mental representations activated by shifting one's attention from one referent to another (topic-shift). The finding that Causal primes pattern like One-Event primes in failing to create a bias for topic-shifting suggests that Causal prime actions are conceptualized – at least at some level – in the same way as One-Event actions. Thus, even though the Causal sequences do involve two sub-events (e.g., event 1: I roll the ball, event 2: the bowling pins fall over), our findings suggest that these sub-events are conceptualized as one (potentially complex) event.

In addition to computing the proportion of continuations with topic-shift vs. topic maintenance, the coherence relations in participants' continuations were also analyzed. They were coded the same way as in Experiment 1. (However, this was not the main aim of Experiment 2: no causal priming was expected in Experiment 2, given that the IC verbs used in this study tend to exhibit a strong bias for explanation relations, which is expected to mask any potential causal priming.) As expected, the most frequent coherence relation in all three conditions was the explanation/because relation (e.g., *Angela frightened Melissa. She was not wearing any makeup.*)[6] All three conditions showed 45–48% explanation continuations (a high proportion, given the large number of different coherence relations that are available), and there were no significant differences between conditions. The overall rates of causal continuations (around 35%), as well as the rate of the consequence and reaction sub-types, also did not differ significantly across conditions. In our opinion, the lack of significant priming for causal (cause + effect) relations is not surprising, given that IC verbs have a strong inherent bias for another kind of continuation (explanation). More generally, when combined with Experiment 1, these patterns suggest that causal priming can be masked in the presence of a stronger discourse-level bias – a finding which fits with the general observation that priming effects (syntactic, semantic, etc.) are often relatively small but nevertheless real.

As a whole, the key finding from Experiment 2 – that referent shifts can be induced by priming with two discrete motor actions – suggests that shifting one's attention from one *event* to another resembles the act of shifting one's attention from one *referent* to another. This points to intriguing similarities between our mental representations of events and entities, something which is also reflected in the fact that they can both be referred to with the same kinds of anaphoric expressions, as illustrated in (5a–b; e.g., Webber, 1991; Kehler and Ward, 2004).

A rollerskate was found behind the old shelves.

< It > was full of cobwebs. (5a)

Peter fell over when rollerskating. < It > was quite a sight! (5b)

[6]When thinking about the subject/object biases and coherence relation biases of IC verbs, it is important to keep in mind that although coherence relations and referential patterns are often related (e.g., with Noun1 IC verbs, explanation continuations tend to start by talking about the preceding subject, Noun1), this is *not* an absolute relationship (see also Pickering and Majid, 2007, p. 784 for related discussion). For example, it is perfectly possible to generate explanation continuations after Noun1 IC verbs that do not start by referring to Noun1 (e.g., *Jason frightened Matt. Matt was very easily startled by the smallest thing.*)

GENERAL DISCUSSION

The two experiments presented in this paper used a cross-modal priming paradigm to investigate how people represent coherence relations in linguistic and non-linguistic domains. Although the coherence relations between sentences play a central role in language comprehension, researchers have come to divergent conclusions about how humans represent and process coherence relations, as well as what the proper taxonomy of coherence relations is. The two experiments in this paper aim to shed some light on these issues, although many questions still remain open for future work.

Experiment 1 explored the domain-generality of coherence relations and the level of detail present in these representations, with a focus on causal relations. Participants carried out different kinds of motor actions (Causal actions, Two-Event actions, and One-Event actions), and provided continuations for agent-patient sentences (e.g., "Mary pinched Kate.") The coherence relations in participants' continuations were analyzed. The results showed that carrying out causal actions – as compared to non-causal actions – made participants more likely to provide causal continuations in the sentence-continuation task. We interpret this as an indication that the mental representations activated by the motor actions overlap, at least in part, with the mental representations used during linguistic discourse-level processing. Furthermore, a detailed analysis of the results shows that the boost in causal continuations is carried by a particular sub-type of causal relations, namely consequence relations (rather than reaction relations). This shows that the mental representations activated by the motor actions contain fine-grained information about the difference between reactions and consequences, and that this is also reflected in the linguistic domain. Although existing models of coherence relations differ in whether they represent coherence relations as logical rules or hierarchical structures (see e.g., Sanders et al., 1993; Webber et al., 2003 for discussion), both of these approaches are in theory compatible with the findings of Experiment 1, as long as they are able to distinguish sub-types of causal relations and allow for some level of representational overlap between discourse-level processing and more domain-general knowledge systems related to causality and event structure.

Because participants were not prevented from encoding the motor actions in linguistic form (e.g., silently describe them), one might wonder about the actual source of the priming effects. To shed light on this, a production study was conducted, and the results indicate that the priming effects observed in Experiment 1 cannot be attributed to syntactic priming. Instead, it seems that the connection is on the level of semantic/conceptual representations: it seems reasonable to conclude that causality representations that were originally triggered by the presentation of non-linguistic, visuo-motor stimuli are tapping into the same (or overlapping) level of representation that is used during the comprehension and production of linguistic stimuli.

To better understand how humans represent coherence relations, we need to gain insights not only into the fine-grained details but also the more general properties of these representations. Experiment 2 explored a truly fundamental property of event sequences, namely the presence of event boundaries. Using the same kind of priming paradigm as in Experiment 1, Experiment 2 looked at the cognitive consequences of shifting

one's attention from one event to another, separate event. The results point to parallels between shifting one's attention from one *event* to another and shifting one's attention from one *referent* to another. More specifically, Two-Event primes were more likely to result in referential shifts in participants' linguistic continuations than One-Event primes. This study also found that Causal primes trigger the same patterns as One-Event primes, suggesting that the two sub-events comprising the Causal sequences are conceptualized as one event (potentially with some kind of internal structure). However, although the outcomes of Experiment 2 shed new light on the nature of discourse-level representations, more work is needed before we can attain a deep understanding of the similarities between shifting between events and shifting between referents, and whether other factors – in addition to event boundaries – may also be influencing the likelihood of topic-shift in these kinds of contexts. Because the research methodology used in this paper (using motor actions as primes for potential discourse-level effects) is still very new, future work will play an important role in helping us to gain a more in-depth understanding of this area.

Broadly speaking, these studies contribute to our understanding of how coherence relations are represented in the mind. The finding that non-linguistic stimuli can influence coherence-related processes in the linguistic domain also fits well with results obtained in earlier work (Kaiser, 2009). In two eye-tracking studies using a priming paradigm, Kaiser (2009) explored how coherence relations presented by means of visuo-spatial/non-linguistic primes or by means of linguistic primes influence pronoun interpretation. Recent research has shown that pronoun interpretation is sensitive to the coherence relations between sentences, as exemplified in ex.(6) where interpretation of "him" is influenced by the coherence relation between the clauses (*causal* vs. *parallel*).

Phil tickled Stanley, and [AS A RESULT] Liz poked him_{Phil}
[him ⇒ Phil]. (6a)

Phil tickled Stanley, and [SIMILARLY] Liz poked $him_{Stanley}$.
[him ⇒ Stanley]. (6b)

Kaiser (2009) conducted two experiments using a paradigm that combines visual-world eye-tracking and priming. In one experiment, participants were presented with visuo-spatial primes, silent video clips that encoded causal relations, similarity relations or other/neutral relations (e.g., Causal = Triangle knocks into circle which falls off a ledge). In another experiment, the coherence relation primes were linguistic (e.g., participants read "The patient pressed the red emergency button near the bed and a nurse quickly ran into the room" for Causal). Participants were then shown a target scene with three characters and heard a sentence like "Phil linded Stanley and Kate hepped him." (Nonce words were used to eliminate verb semantics). Participants' eye-movements were used to assess which entities were considered as referents for "him." The results show that pronoun interpretation can indeed be primed by coherence relations in preceding linguistic and visual input, even when primes and targets are connected only on the level of abstract coherence relations, and even when primes are presented in a non-linguistic modality.

These earlier findings, like the two productions studies presented in the current paper, highlight the important role that coherence relations play in language processing, and suggest that coherence relations make reference to domain-general representations. Broadly speaking, this line of research has the potential to offer new insights into the nature of the interface between linguistic and non-linguistic representations. It taps into one of the central questions in psycholinguistics, namely the extent to which language is distinct from other cognitive processes vs. supported by domain-general processes.

This research was partially supported by funding from the University of Southern California Undergraduate Research Associates Program and NIH grant 1R01HD061457 to Elsi Kaiser. We wish to thank Noelle Garza, Alexandra Ruelas, and Boutaina Cherqaoui for assistance with the experiments and data analysis. Thanks are also due to the two reviewers who provided extremely useful and insightful comments.

REFERENCES

Arai, M., Van Gompel, R. P. G., and Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cogn. Psychol.* 54, 218–250.

Asher, N., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Baayen, H. R., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412.

Bock, K. (1986). Syntactic persistence in language production. *Cogn. Psychol.* 18, 355–387.

Bock, K., and Griffin, Z. (2000). The persistence of structural priming: transient activation or implicit learning? *J. Exp. Psychol. Gen.* 129, 177–192.

Branigan, H. P., Pickering, M. J., and Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118, 172–189.

Briner, S., Virtue, S., and Kurby, C. (2012). Processing causality in narrative events: temporal order matters. *Discourse Process.* 49, 61–77.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., and Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: a combined TMS and behavioral study. *Brain Res. Cogn. Brain Res.* 24, 355–363.

Casasanto, D., and Dijkstra, K. (2010). Motor action and emotional memory. *Cognition* 115, 179–185.

Chafe, W. L. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Garvey, C., and Caramazza, A. (1974). Implicit causality in verbs. *Linguist. Inq.* 5, 459–464.

Gati, I., and Tversky, A. (1984). Weighting common and distinctive features in conceptual and perceptual judgments. *Cogn. Psychol.* 16, 341–370.

Gentner, D., and Markman, A. B. (1997). Structure mapping in analogy and similarity. *Am. Psychol.* 52, 45–56.

Glenberg, A., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., and Buccino, G. (2008). Processing abstract language modulates motor system activity. *Q. J. Exp. Psychol.* 61, 905–919.

Glenberg, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Hobbs, J. (1979). Coherence and coreference. *Cogn. Sci.* 3, 67–90.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446.

Johnston, J., and Welsh, E. (2000). Comprehension of "because" and "so": the role of prior event representation. *First Lang.* 20, 291–304.

Kaiser, E. (2009). "Effects of anaphoric dependencies and semantic representations on pronoun interpretation," in *Anaphora Processing and Applications*, eds S. L. Devi, A. Branco, and R. Mitkov (Heidelberg: Springer), 121–130.

Kanizsa, G., and Vicario, G. (1968). "The perception of intentional reaction," in *Experimental Research on Perception*, eds G. Kanizsa

and G.Vicario (Trieste: University of Trieste), 71–126.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. Stanford: CSLI Publications.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. (2008). Coherence and coreference revisited. *J. Semant.* 25, 1–44.

Kehler, A., and Ward, G. (2004). "Constraints on ellipsis and event reference," in *Handbook of Pragmatics*, eds L. R. Horn and G. Ward (Oxford: Blackwell), 383–403.

Knott, A. (1993). "Using cue phrases to determine a set of rhetorical relations," in *Intentionality and Structure in Discourse Relations: Proceedings of the ACL SIGGEN Workshop*, ed. O Rambow, Columbus: Ohio State University.

Koornneef, A. W., and Van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: evidence from self-paced reading and eye tracking. *J. Mem. Lang.* 54, 445–465.

Lambrecht, K. (1994). *Information Structure and Sentence form: TOPIC, Focus, and the Mental Representation of Discourse Referents.* Cambridge: Cambridge University Press.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago, IL: University of Chicago Press.

Mann, W. C., and Thompson, S. A. (1986). Relational propositions in discourse. *Discourse Process.* 9, 57–90.

Mann, W. C., and Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text* 8, 243–281.

Michotte, A. (1946/1963). *The Perception of Causality*, trans. T. R. Miles and E. Miles (New York: Basic Books).

Noordman, L. G. M. (2001). On the production of causal-contrastive "although" sentences in context," in *Text Representation: Linguistic and Psycholinguistic Aspects*, eds T. Sanders, J. Schilperoord, and W. Spooren (Amsterdam: John Benjamins), 153–180.

Pickering, M. J., and Branigan, H. P. (1998). The representation of verbs: evidence from syntactic persistence in written language production. *J. Mem. Lang.* 39, 633–651.

Pickering, M. J., and Majid, A. (2007). What are implicit causality and consequentiality? *Lang. Cogn. Process.* 22, 780–788.

Pulvermüller, F., Shtyrov, Y., and Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition: an MEG study using the mismatch negativity. *J. Cogn. Neurosci.* 17, 884–892.

Reinhart, T. (1982). Pragmatics and linguistics: an analysis of sentence topics. *Philosphica* 27, 53–94.

Rohde, H. (2008) *Coherence-Driven Effects in Sentence and Discourse Processing.* Ph.D. Dissertation, University of California, San Diego.

Sanders, T. J. M. (2005). "Coherence, causality and cognitive complexity in discourse," in *Proceedings of the First International Symposium on the Exploration and Modelling of Meaning*, eds M. Aurnague and M. Bras (Toulouse: Universite de Toulouse-le-Mirail), 31–46.

Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Process.* 15, 1–35.

Sanders, T., Spooren, W., and Noordman, L. (1993). Coherence relations in a cognitive theory of discourse representation. *Cogn. Linguist.* 4, 93–133.

Schlottmann, A., Ray, E., Mitchell, A., and Demetriou, N. (2006). Perceived physical and social causality in animated motions: spontaneous reports and ratings. *Acta Psychol. (Amst.)* 123, 112–143.

Simmons, S., and Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition* 108, 781–795.

Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Lang. Cogn. Process.* 9, 519–548.

Stewart, A. J., Pickering, M. J., and Sanford, A. J. (2000). The time course of the influence of implicit causality information: focusing versus integration accounts. *J. Mem. Lang.* 42, 423–443.

Stukker, N., Sanders, T., and Verhagen, A. (2008). Causality in verbs and in discourse connectives. Converging evidence of cross-level parallels in Dutch linguistic categorization. *J. Pragmat.* 40, 1296–1322.

Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *J. Exp. Psychol. Gen.* 138, 236–257.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J. Cogn. Neurosci.* 17, 273–281.

Trabasso, T., and van den Broek, P. (1985). Causal thinking and the representation of narrative events. *J. Mem. Lang.* 24, 912–630.

van den Broek, P. W. (1990). "Causal inferences in the comprehension of narrative texts," in *Psychology of Learning and Motivation: Inferences and Text Comprehension*, eds A. C. Graesser and G. H. Bower (New York, NY: Academic Press), 175–196.

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Lang. Cogn. Process.* 6, 107–135.

Webber, B. L., Stone, M., Joshi, A. K., and Knott, A. (2003). Anaphora and discourse structure. *Comput. Linguist.* 29, 545–587.

Wolfe, M. B. W., Magliano, J. P., and Larsen, B. (2005). Causal and semantic relatedness in discourse understanding and representation. *Discourse Process.* 39, 165–187.

Zacks, J. M., and Swallow, K. M. (2007). Event segmentation. *Curr. Dir. Psychol. Sci.* 16, 80–84.

Zwaan, R. A., and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychol. Bull.* 123, 162–185.

Zwaan, R. A., and Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *J. Exp. Psychol. Gen.* 135, 1–11.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 October 2011; accepted: 30 April 2012; published online: 11 June 2012.

*Citation: Kaiser E (2012) Taking action: a cross-modal investigation of discourse-level representations. Front. Psychology **3**:156. doi: 10.3389/fpsyg.2012.00156*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 21 November 2011
doi: 10.3389/fpsyg.2011.00340

Fast mapping of novel word forms traced neurophysiologically

Yury Shtyrov [1,2] *

[1] Cognition and Brain Sciences Unit, Medical Research Council, Cambridge, UK
[2] Cognitive Brain Research Unit, Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Andriy Myachykov, University of Glasgow, UK
Mikael Roll, Lund University, Sweden
Kambiz Tavabi, Children's Hospital of Philadelphia, USA

***Correspondence:**
Yury Shtyrov, Cognition and Brain Sciences Unit, Medical Research Council, 15 Chaucer Road, CB2 7EF Cambridge, UK.
e-mail: yury.shtyrov@mrc-cbu.cam.ac.uk

Human capacity to quickly learn new words, critical for our ability to communicate using language, is well-known from behavioral studies and observations, but its neural underpinnings remain unclear. In this study, we have used event-related potentials to record brain activity to novel spoken word forms as they are being learnt by the human nervous system through passive auditory exposure. We found that the brain response dynamics change dramatically within the short (20 min) exposure session: as the subjects become familiarized with the novel word forms, the early (~100 ms) fronto-central activity they elicit increases in magnitude and becomes similar to that of known real words. At the same time, acoustically similar real words used as control stimuli show a relatively stable response throughout the recording session; these differences between the stimulus groups are confirmed using both factorial and linear regression analyses. Furthermore, acoustically matched novel non-speech stimuli do not demonstrate similar response increase, suggesting neural specificity of this rapid learning phenomenon to linguistic stimuli. Left-lateralized perisylvian cortical networks appear to be underlying such fast mapping of novel word forms unto the brain's mental lexicon.

Keywords: brain, cortex, language, word, event-related potential, electroencephalography, lexical memory trace, fast mapping

INTRODUCTION

As a communication tool, human language is far more complex than any signaling system developed by other animal species. Amongst the many features making human language unique is the impressive size of our vocabularies, which reach into tens of thousands of words (Corballis, 2009). To acquire this knowledge, humans learn new words with high speed and efficiency – as children acquiring their native tongue and as adults mastering a new one. This capacity for rapid learning of language, also known as "fast mapping," has been demonstrated in numerous behavioral studies and observations (Carey and Bartlett, 1978; Dollaghan, 1985) which have indicated immediate behavioral effects of fast word learning present even before the nervous system has had a chance of consolidating the new information. However, the neural underpinnings of this crucial human skill still remain obscure. On the systems level, much experimentation has been done on longer-term effects of learning revealing neural correlates of days and weeks of practice or at least an overnight consolidation (see Davis and Gaskell, 2009, for a review), whereas the rapid aspect of word learning has remained a difficult task for neurobiological studies.

Indeed, addressing immediate plastic changes in the healthy human brain, as it is learning new words, is not a trivial task. Unlike animal research, invasive measures that provide direct assessment of neural activity are generally not possible in humans. This implies the need to use other tools that either address neural activity indirectly (such as behavioral or hemodynamic methods) or, even if they deal with mass neuronal activation (such as electro and magnetoencephalography, EEG/MEG), their limited resolution normally requires presentation of multiple trials to acquire a stable image of brain activity. These methodological limitations prevent straightforward recording of dynamic neural changes in the learning process. This is why most neuroimaging attempts so far could only provide a derived and abstracted picture of fast learning processes in the brain, failing to capture the online progression of language elements from novel to learnt. To date, only a small number of experiments combining modern neuroimaging tools with carefully designed linguistic paradigms have been preformed to explore the human brain dynamics in language learning.

One such study trained adult functional magnetic resonance imaging (fMRI) subjects on a novel vocabulary of concrete nouns that were assigned meaning via a word–picture associative learning paradigm, which took place during the scanning (Breitenstein et al., 2005). Rather than comparing different conditions, this study monitored changes in the hemodynamic brain activation throughout the experiment by quantifying BOLD responses over five consecutive experimental sub-blocks. It showed changes in the hippocampus in the learning exposure accompanied by a complex pattern of activity involving a variety of neocortical structures: selective activation of right inferior-frontal gyrus, suppression in left fusiform gyrus, and activation increase in left inferior parietal lobe. Investigations using positron-emission tomography (PET) showed that changes in activity in bilateral posterior superior temporal gyri correlate with behavioral performance in non-word learning task (Majerus et al., 2005). Another PET study indicated

a left-lateralized network of neocortical areas – temporal lobe, inferior-frontal gyrus, temporo-parietal junction – as taking part in rapid word learning, along with parahippocampal structures (Paulesu et al., 2009). Importantly, such studies not only confirm hippocampal involvement in encoding that had been known from previous animal neurophysiology research and neuropsychological studies in brain-damaged patients, but they also indicate a complex neocortical pattern of activation and de-activation that takes place in the learning process. On one hand, this does map onto a generally accepted two-stage or "complementary" learning systems approach, which maintains that initial encoding takes place in hippocampus with a later slow-rate (days/weeks) transfer of memory representations to neocortex (McClelland et al., 1995); on the other hand, this questions the slowness of neocortical memory trace formation and clearly suggests neocortical involvement in initial encoding stages.

Whilst hemodynamic brain imaging has exquisite spatial resolution, its temporal resolution – on the order of seconds – is poor; furthermore, it does not measure neural processes directly but addresses them by proxy, via cerebral blood flow and metabolism. For these principled reasons, metabolic neuroimaging cannot measure rapid neuronal activations that are known to take place on the millisecond range. Language-elicited brain dynamics is known to unfold extremely rapidly with a number of processing stages reflected in complex neuronal activation patterns in the first few hundred milliseconds of stimulus arrival (Friederici, 2002; Pulvermüller and Shtyrov, 2009; Shtyrov et al., 2010a). Clearly, to better understand neural processes of language learning, there is a need for a more direct measure of electric neuronal activity; this can be afforded by neurophysiological time-resolved imaging tools such as electroencephalography.

To explore electrophysiological correlates of rapid word learning, some EEG studies have used N400, a negative deflection in the brain's event-related potentials that is known to be sensitive to lexical and semantic stimulus features. Mestres-Misse et al. (2007), whose subjects were required to discover the meaning of a visually presented novel word from its context, found that just after a few exposures to novel words, their N400 response amplitudes were virtually indistinguishable from those to previously known words. Very similar electrophysiological dynamics was obtained in a more recent N400 study using context-restricted novel word learning, also in the visual modality (Borovsky et al., 2010). Interestingly, in an EEG study that involved learning an artificial language, an increase of N400 in response to newly learnt words was found already after 1 min of exposure (De Diego Balaguer et al., 2007).

Whilst N400 is an established linguistic ERP component, in sentential context it likely reflects not only, and not so much the word learning processes *per se*, but rather the integration of the new items into a larger context (Friederici, 2002). It has also been argued that neural access to lexical word information commences much earlier than 400 ms and can already be reflected in evoked responses at 100–150 ms (Shtyrov et al., 2005; Shtyrov and Pulvermüller, 2007; Pulvermüller et al., 2009). Thus, the need to directly address learning of individual words as such is still open. Behavioral studies suggested that a mere repetitive exposure to a novel word form creates a lexical entry (Gaskell and Dumay, 2003). This was directly tested in a recent EEG study (Shtyrov et al., 2010b), where the subjects were passively exposed in a very short session to a repetitive presentation of the same novel word form, with an acoustically similar real word serving as a control. Importantly, whilst the N400 studies above used visual presentation, this experiment was performed in the auditory modality, the native modality for language in which most of natural language acquisition occurs in real life. To test the dynamics of the stimuli's lexical status in the subjects' mental lexicon, this study used passive oddball stimulus presentation that is known to generate diverging patterns for words and unfamiliar pseudo-words: the early (∼120 ms) passive oddball response to a spoken word is enhanced in comparison with similar pseudo-word, and this enhancement is believed to be a neural signature of a word-specific memory trace activation (Pulvermüller and Shtyrov, 2006; Shtyrov et al., 2010a). In the first minutes of the exposure session, an enhanced activity for known words was found, indexing the ignition of their underlying memory traces. However, just after ∼14 min of learning exposure, the novel word forms exhibited a significant increase in response magnitude matching in size with that to real words. This activation increase, as it was proposed, reflects rapid mapping of new word forms onto neural representations formed in left temporal/perisylvian neocortex.

This study was, however, limited in its findings as it only used a single token of novel word form. This was presented in an oddball paradigm, a rather unnatural stimulus presentation mode in which one frequent stimulus is presented hundreds of times and is occasionally replaced by a diverging auditory event. Although the single-item approach is similar to the earliest behavioral research which reported fast mapping of novel words using a single token (Carey and Bartlett, 1978) and such findings cannot be refuted *per se*, generalizability of such a result is rather limited. Furthermore, none of the previous studies controlled the specificity of fast mapping effects to language by employing comparable non-linguistic conditions. In this study, we have set out to overcome the shortcomings of earlier research. We investigated online neural correlates of novel word form learning using a small acoustically matched group of known words and novel spoken word forms which were presented, at a natural speech rate, to experimental participants in a passive auditory exposure together with acoustically matched novel non-speech stimuli, whilst online measures of the participants' brain activity were taken using multi-channel electroencephalographic recordings.

MATERIALS AND METHODS

SUBJECTS

Sixteen healthy right-handed (handedness assessed according to Oldfield, 1971) native Finnish-speaking subjects (Helsinki University students, age 18–29, seven males) with normal hearing and no record of neurological diseases were presented with spoken Finnish language stimuli in two experimental conditions. All subjects gave their written consent to take part in the study and were paid for their participation.

AUDITORY STIMULATION

For stimulus presentation, we employed a small group of controlled bi-syllabic stimuli which were closely matched in their acoustic features and were produced by recombining the same set

of two first and four second syllables to generate eight spoken items with different lexical properties: four previously unfamiliar novel word forms (so called "pseudo-words") and four known words used as a control, as well as two additional non-speech controls. Two Finnish syllables [pa] and [ta] were combined with syllables [ko], [ku], [ke], [ki], which resulted in the following combinations: *pakko*, **pakku*, *pakki*, **pakke*, in one of the conditions, and **takko*, *takku*, *takki*, **takke* in the other condition (double consonant in Finnish stands for a geminate stop signifying the extended silent closure before the [k], 275 ms in this case; pseudo-words are preceded with an asterisk). Note that the stimulus combinations were minimally different in their acoustic features with the final consonant–vowel transition being sufficient to identify each item *per se* as well as differentiate between the known words and novel pseudo-words. This made sure that the time point when any possible lexical effects could commence was the same across all stimuli of interest – at the onset of the second syllable. This is essential for analyzing auditory ERP recordings that are highly sensitive to temporal and other physical-acoustic features of the stimuli; in this design, we could time-lock responses to the same time point for all stimuli. These minimal word-final differences also meant that the stimuli within each block belonged to the same cohort, i.e., had common lexical neighbors with similar onsets (as *ta*- and *pa*-starting stimuli were presented in two separate blocks). Effectively, the range of possible alternatives was restricted by the experimental settings to the stimulus set as no other completions were possible in each experimental block.

For stimulus production, we recorded multiple repetitions of these syllables uttered by a female native speaker of Finnish and selected a combination of the six items whose vowels matched in their fundamental frequency (F0) as well as sound energy and overall duration (**Figure 1**). The sounds were normalized to have the same loudness by matching their root-mean-square (RMS) power; this was separately normalized for the first ([pa]/[ka]) and for the second ("word-final") syllables. Further, a signal-correlated noise (SCN) was produced by subjecting acoustic white noise to a fast Fourier-transform (FFT) filter, whose profile was modeled after the actual second syllables; the filtered noise was then given a temporal envelope of a CV-syllable and combined with the same two first syllables to produce two non-speech control stimuli. All individual syllables (including non-speech SCN) were 100 ms long and all complete stimuli were 475 ms in duration. The stress was always placed on the first syllable, as it is standard in the Finnish language. For the analysis and production of the stimuli we used the Cool Edit 2000 program (Syntrillium Software Corp., AZ, USA).

Given previous behavioral linguistic research indicating that word learning reaches a plateau at ∼150 repetitions in a short behavioral exposure (Pittman, 2008), we presented our experimental subjects with the novel spoken pseudo-words, control words, and SCN stimuli 160 times per each stimulus in a passive listening task lasting approximately 20 min. Each of the two blocks ([pa]/[ta]) included 160 pseudo-random repetitions of five (four speech and one SCN) stimuli. All stimuli were presented via headphones at 50 dB above individual hearing threshold. Stimulus onset asynchrony was 750 ms, approximating natural speech rate in Finnish (Valo, 1994). The order of the two blocks was counterbalanced across the subject group. Previous research has

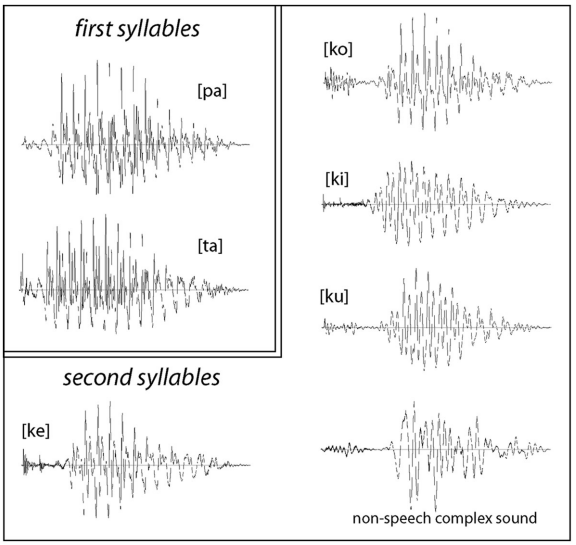


FIGURE 1 | Waveforms of acoustic stimuli used in the experiments: all stimuli were composed of the same first syllables [pa] and [ta], which were recombined (after a 275 silent closure) with the second syllables [ku], [ko], [ke] [ki], and a matched non-speech sound. The stimuli were maximally matched for their acoustic properties, whilst their lexical status as familiar or novel items was systematically modulated.

suggested that initial lexical processing is automatic and that early neurophysiological effects may be masked by focused attention (Garagnani et al., 2009; Shtyrov et al., 2010a); participants' attention was therefore diverted from the stimuli to a silent video film of their own choice whilst they listened passively to the auditory stimuli, as it was done in a previous study that successfully traced formation of novel memory traces for single words (Shtyrov et al., 2010b).

ELECTROENCEPHALOGRAPHIC RECORDING

Subjects were seated in an electrically and acoustically shielded chamber. During the stimulation, electric activity of the subjects' brain was continuously recorded (passband 0.01–100 Hz, sampling rate 500 Hz) with a 64-channel EEG set-up (Compumedics Neuroscan, El Paso, TX, USA), using gold-plated Ag/AgCl electrodes mounted in an extended 10–20-system custom-made electrode cap (Virtanen et al., 1996) and a separate nose reference electrode. To control for eye-movement artifacts, horizontal and vertical eye movements were recorded using two bipolar electrooculogram (EOG) electrodes.

EEG DATA PROCESSING

The recordings were later filtered off-line (passband 1–20 Hz, 12 dB/oct). Event-related potentials were obtained by averaging epochs, which started 50 ms before the stimulus disambiguation point (second syllable onset) and ended 400 ms thereafter; −50 to 0 ms interval was used as a baseline. Epochs with voltage variation exceeding 100 μV at any EEG channel or at either of the two EOG electrodes were discarded; on average, this led to 117 accepted trials for each stimulus type. The remaining EEG data were recomputed

against average reference. Following this, three types of analysis were used. We first compared data subsets covering the initial and final 10% of the learning session. Notably, these amounted to 16 or fewer trials for each individual stimulus, which is substantially below the standard auditory ERP studies that typically use in excess of 100 trials for averaging; as we hypothesized that rapid learning could occur within a short time interval, we had to limit the number of trials to see any potential learning effects. To overcome the low signal-to-noise ratio (SNR) resulting from the inherent small number of trials, we pulled together data from all novel pseudo-words and, separately, known words. Based on previous research (Shtyrov et al., 2010a,b), we extracted data from fronto-central midline electrodes where the auditory evoked response is typically maximal (Fz, FCz) in an *a priori* defined 20-ms window at 110–130 ms and submitted these to analyses of variance (ANOVA) with the factors Stimulus type (Word/Pseudo-word) and Exposure time (early/late in the session). As visual inspection of responses showed an additional presence of an earlier peak (∼80 ms), a second 20 ms time window centered on this earlier deflection was added to the analyses *post hoc*.

Our second analysis, aimed at finer-scale temporal changes in the responses over the course of the session, applied linear regression on individual subjects' peak amplitude data obtained from consecutive 10% intervals for both word and pseudo-word responses. Having fitted the least-squares line to individual amplitude measurements for each subject, we submitted regression coefficients to ANOVAs in order to verify significance of any observed differences between stimulus types. Brain Vision Analyzer 1.05 (Brain Products, Gilching, Germany) was used for processing the EEG signal, Matlab 7.0 programming environment (Mathworks, Natick, MA, USA) was used for in the linear regression analyses; statistical analysis was implemented in Matlab 7.0 and in Statistica 7.1 (Statsoft, Tulsa, OK, USA).

In the final analysis, aimed at localizing cortical sources of the found learning effect (response increase for the novel pseudo-word), we performed L2 minimum-norm current estimation on ERP difference between the pseudo-word trials collected in the end and start (10%, i.e., last vs. first 2 min) of the exposure block. This distributed source analysis does not make *a priori* assumptions about underlying generators and attempts to minimize the overall activity that can account for the recorded electric potentials (Ilmoniemi, 1993). MNE solutions were calculated for grand-average responses rather than individual data; calculating solutions on grand-average data has a benefit of substantially reduced noise and therefore improved SNR which MNE solutions are highly sensitive to (hence individual source solutions were not possible here due to the low SNR inherent to the small number of trials under consideration), although prevents assessing results statistically. A three-layer boundary element model with triangularized gray matter surface of a standardized brain (Montreal Neurological Institute) was used for computing source reconstruction solutions. The solutions were restricted to smoothed gray matter surface. CURRY 6.1 software (Compumedics Neuroscan, Hamburg, Germany) was used for these procedures. Based on the previous studies, our expectation was that of left-lateralized perisylvian activation for the newly formed memory representations.

RESULTS

All items elicited evoked responses, and ERPs were successfully calculated for the word and pseudo-word stimuli both early and late in the exposure session (**Figures 2** and **3**). Within a short time after the divergence point (∼70–130 ms), the ERP temporal dynamics demonstrated differences for the novel and familiar items early and late in the exposure session. The first analysis, concentrated on the *a priori* defined window centered on 120 ms, indicated a fronto-central maximum of positive polarity that showed a significant interaction Stimulus type × Exposure time [$F(1,15) = 13.45$, $p = 0.0023$]. Investigating this interaction with planned comparisons, we found that it was due to the word response remaining unchanged between the start and the end of the exposure block ($p > 0.5$), while the pseudo-word response enhanced significantly with time [$F(1,15) = 16.79$, $p = 0.0009$]. Visual inspection of the data (**Figure 2**) indicated that exposure-related ERP effects were occurring also in an earlier time window, with a word-elicited maximum peaking at 80 ms. To account for this earlier activation, we added a second 20-ms window (70–90 ms) to the analysis. This combined analysis supported the Stimulus type × Exposure time interaction [$F(1,15) = 5.83$, $p = 0.0289$]; again, planned comparisons confirmed that it was due to the absence of changes in the word response ($p > 0.9$) and a significant increase in the pseudo-word activity [$F(1,15) = 11.62$, $p = 0.0034$]. A marginally significant interaction of the newly introduced factor Window (80 vs. 120 ms) with Stimulus type [$F(1,15) = 4.03$, $p = 0.06$] suggested an earlier peak for the word than pseudo-word stimuli (also visible in the ERP patterns). We therefore directly compared the slightly later activation for pseudo-words with the earlier word peak. This comparison, for the third time, confirmed the differential word/pseudo-word dynamics over the learning session as a significant interaction [$F(1,15) = 11.73$, $p = 0.0038$]. Furthermore, investigation of this interaction with planned comparisons showed that whilst the word response significantly exceeded that to pseudo-word in the beginning of the session [$F(1,15) = 6.10$, $p = 0.025$], the difference between the two was absent in the end of the exposure ($p > 0.13$).

To quantify the development of language-evoked brain activity throughout the entire recording session, linear regression analysis was applied to word- and pseudo-word-elicited activation calculated for successive sub-averages (10%) obtained from each individual, pulled across both analysis windows (**Figure 4**). Least-squares lines fitted to word ERPs demonstrated a stable pattern, whereas for the newly learnt pseudo-words the regression analysis showed a significant increase in event-related activity with exposure time. The specific increase of brain responses to pseudo-words was further confirmed by a statistical comparison of regression slopes (beta values) obtained from each subject individually and entered into group analysis [$F(1,15) = 4.89$; $p < 0.045$].

ERP topography (**Figure 2**) suggested that the word responses had a consistent bias toward left-hemispheric lateralization early and late in the training session, whilst the pseudo-word response appeared to shift from a central to a left-biased distribution with exposure progress (see also maps in **Figure 3**); this interaction, however, did not reach significance. To further localize the cortical sources potentially underlying the rapid emergence of memory

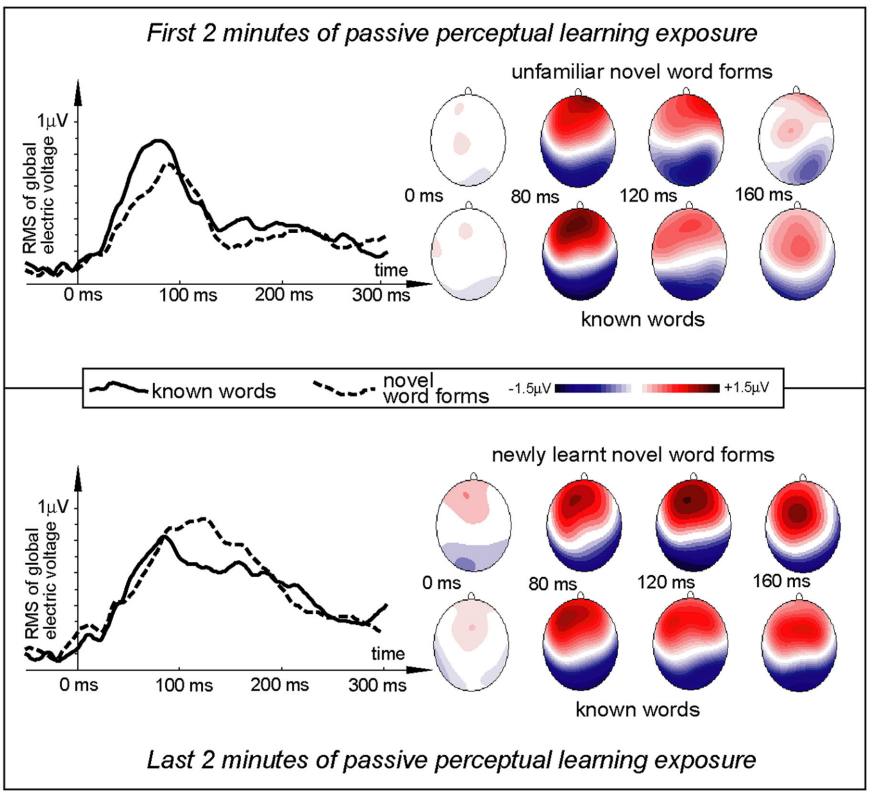


FIGURE 2 | Electric brain response (global activation computed as RMS across all EEG electrodes; grand-average data) for word and pseudo-word stimuli early and late in the learning session. Responses are time-locked to the stimulus uniqueness points (second syllable onsets) when each stimulus could first be identified. Note the larger word response early in the session and the pseudo-word response increase by the end of the exposure.

traces for novel word forms, L2 minimum-norm current estimation was applied to ERP difference between the pseudo-word trials collected in the end and start of session. Sources of this neurophysiological effect were localized to bilateral temporal and inferior-frontal cortices with a noticeable lateralization of activity to left-perisylvian neocortex (**Figure 5**), in line with the ERP signal topography (**Figure 3**) and our original predictions. As grand-average data were used in this analysis in order to improve the SNR for computing the solutions, these results could not be verified statistically and should therefore be treated with caution.

Finally, the non-speech SCN stimulus did not exhibit any significant changes over the duration of repetitive perceptual exposure. Its time course (**Figure 6**) was markedly different from that elicited by the spoken stimuli and in the early interval near 100 ms was suggestive of a response decline with the reverse taking place after 200 ms. However, no significant exposure-related differences could be located ($p > 0.6$).

DISCUSSION

We recorded brain's responses to previously unfamiliar novel spoken word forms, acoustically matched real familiar words and non-linguistic sounds. These were randomly and repetitively presented in a passive auditory exposure that lasted approximately 20 min. Electric brain responses were generated by all types of stimuli; changes in their dynamics over the course of the perceptual learning session were scrutinized using a factorial analysis which compared ERPs in the beginning and end of the recording, and a linear regression approach that looked for stable patterns over successive sub-averages throughout the session.

The earliest activity that was registered here and exhibited differential dynamics was that around 70–130 ms from the point in time when the information in the auditory input allowed for stimulus identification. This deflection had a fronto-central distribution of positive polarity (using average reference) and showed a markedly different dynamics between the stimulus types. The familiar known words produced a stable pattern with minimal changes between the beginning and the end of the session. This stability is in line with previously postulated robustness of neural circuits acting as word-specific memory traces (Garagnani et al., 2009; Shtyrov, 2010). In contrast, novel word forms, which initially produced a smaller response than that to words, demonstrated a dramatic change with the exposure progress and finally matched in size (and visually even overtook) the response to words.

This pseudo-word-specific activation modulation with exposure time, as we would like to propose, reflects rapid mapping of new word forms onto neural representations. Importantly, this activation is remarkably early (∼100 ms) and occurs in a passive perceptual exposure, when the subjects are not paying attention to

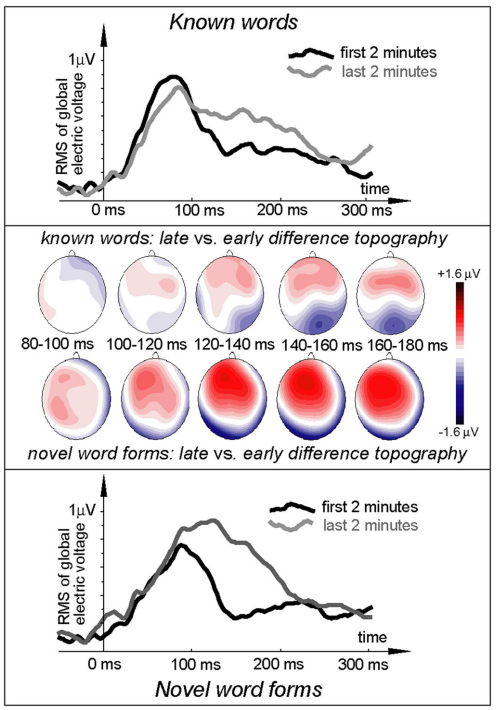


FIGURE 3 | Electric brain response (global activation computed as RMS across all EEG electrodes; grand-average data) for word and pseudo-word stimuli early and late in the learning session and voltage topography maps for comparison between the early and late response (based on "late" minus "early" subtraction). Note the larger change in the pseudo-word response by the end of the exposure, topographically visible as an increased left-frontal positivity in the voltage maps.

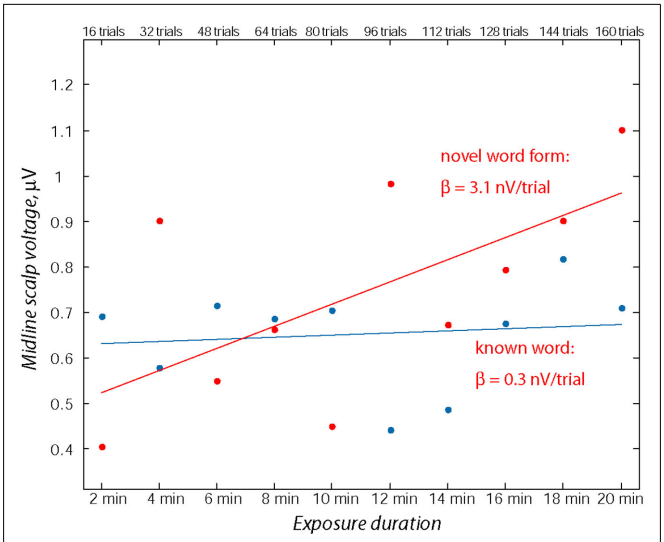


FIGURE 4 | Assessment of ERP magnitude change through the exposure session using linear regression over consecutive 10% sub-blocks. Note the relative stability of the word response in contrast with the marked increase in the pseudo-word response amplitude. Data from both time windows (70–90 and 110–130 ms) from midline electrodes (Fz, FCz) were used for computing linear regression for each participant's responses to known words and novel pseudo-words.

the stimuli. These two factors largely exclude the possibility that it may be linked to secondary post-comprehension processes, an argument that could in principle be made in relation to metabolic or even N400 studies. Such a neural correlate of rapid word form learning emerging within minutes of passive perceptual exposure confirms that our brain may effectively form new linguistic memory circuits online, as it gets exposed to novel speech patterns in the sensory input.

A similar result of a rapidly increased activity for a novel pseudo-word has been demonstrated earlier (Shtyrov et al., 2010b). However, the important advance in the current study is that it used multiple tokens of word and pseudo-word stimuli presented within the natural range of speech rate, thus offering a much stronger experimental base for this phenomenon. Furthermore, here we have also employed a non-speech control stimulus set. Although the stimuli it included were highly similar acoustically to the speech syllables, they generated a different ERP dynamics in general and, most importantly, did not exhibit any learning-related changes. The latter suggests that although the human capacity to rapidly learn new words may have common roots with animal learning mechanisms (Kaminski et al., 2004), it appears to have developed into a sophisticated neural machinery specific to language learning. Even if rapid learning is not specific to human language function (as it has been argued by, e.g., Markson and Bloom, 1997) and may be an expression of a more general neurobiological learning mechanism, the extremely efficient application of this mechanism to the learning of vocabularies of thousands of words is, of course, a human feature. This feature is potentially facilitated by human-specific neuroanatomical advantages in the form of efficient connections within left temporo-frontal perisylvian networks (Catani et al., 2005; Saur et al., 2008).

Indeed, left-hemispheric temporo-frontal structures were indicated as playing the dominant part in the rapid learning of novel words in the current study. Although our source analysis here was based on grand-average data and thus not verifiable statistically, these structures were also indicated by previous metabolic imaging studies of fast mapping (Majerus et al., 2005; Rauschecker et al., 2008; Paulesu et al., 2009). The brain structures engaged by such rapid passive word form learning are part of those also effective in the processing of meaningful words, such as superior temporal cortex included in the "what" stream of auditory processing (Rauschecker and Scott, 2009). Partial involvement of the right hemisphere that is suggested by the source analysis here has also been shown before, specifically a strong involvement of right inferior-frontal gyrus in fast mapping of novel words as seen in fMRI (Breitenstein et al., 2005) is confirmed by the current source analysis results. Importantly, the present study along with the earlier studies we have reviewed above makes a strong case for a network of neocortical areas that take part in online word acquisition and that may include most notably perisylvian structures of the left hemisphere (temporal lobe, inferior-frontal gyrus), as well as temporo-parietal, premotor, and prefrontal regions. This network may be underpinning a neocortical "fast track" for

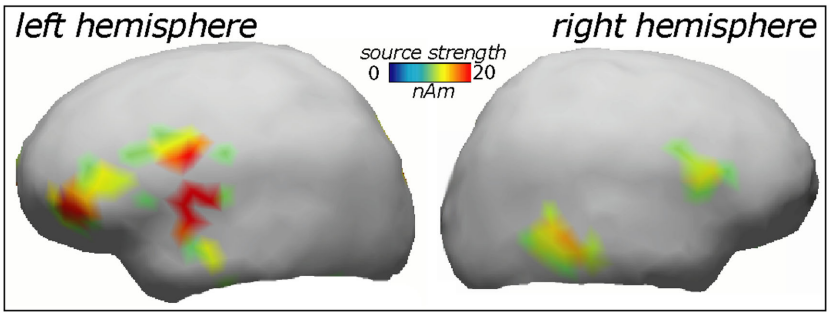


FIGURE 5 | Cortical source distributions (L2 minimum-norm) in the left and right cerebral hemisphere accounting for the increase in novel word form activation over the exposure session.

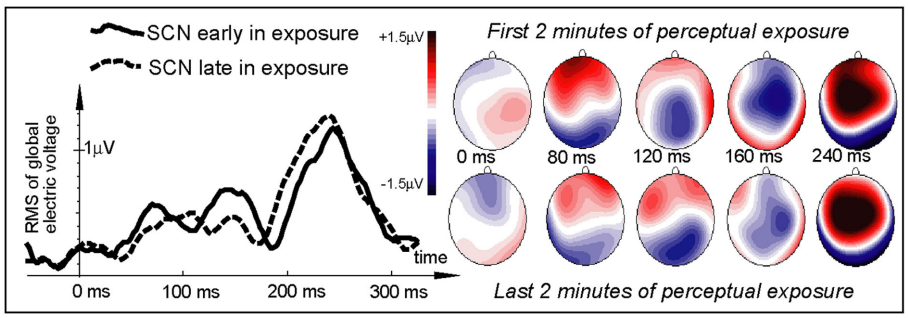


FIGURE 6 | Electric brain response (global activation computed as RMS across all EEG electrodes; grand-average data) for the non-speech signal-correlated noise control stimuli early and late in the learning session. Note the marked difference in the SCN time course from that elicited by the spoken stimuli (cf. **Figure 3**). No significant exposure-related differences could be located for this non-speech elicited activation.

word acquisition which subserves the vital function of rapid language learning not directly dependant on long-term consolidation processes traditionally linked to hippocampus (McClelland et al., 1995; Born et al., 2006). This suggestion is well supported by a recent neuropsychological investigation showing a near-normal fast mapping ability in patients with severely damaged hippocampus that critically depends on intact left temporal cortex (Sharon et al., 2011).

In addition to supporting the previously made notion of rapid (∼100 ms) lexical effects in auditory ERPs that can also be used for tracking word memory trace formation, this study has shown three noticeable differences from the earlier investigations. First, in at least one previous similar study that demonstrated such an effect, it had a negative surface polarity (Shtyrov et al., 2010b), whereas here the entire action is occurring on the positive end of the voltage scale, although the fronto-central distribution largely remains the same. This is likely explained by differences in the paradigm we employed: whilst the previous investigation used an oddball single token approach and monosyllabic stimuli, here were presented a selection of different bi-syllabic items mixed equiprobably. The higher (and more natural) rate of stimulus presentation here, along with the analysis focus on the second syllables may mean that the negativity usually seen at this latency is greatly suppressed due to habituation resultant from continuous auditory stimulation (Rosburg et al., 2006). In time, the effects seem to generally correspond to the traditional N100 latency range as well as the time when lexical MMN effects have been demonstrated, and could thus be related to these auditory ERPs; however, the unusual polarity dynamics call for future exploration of these effects' neural origins. Interestingly, in at least one earlier EEG experiment on rapid language learning, an increase in frontal positivity with peak latency shortly before 200 ms (i.e., P2 range) has also been observed, but it was linked to rule acquisition rather than word learning processes (De Diego Balaguer et al., 2007).

Second, the results suggested a later peak for the pseudo-word response (particularly noticeable in the end of the learning exposure, **Figures 2** and **3**) than for the word-elicited ERP. Although this difference was only marginally supported by statistics ($p = 0.06$), it indicates a potentially interesting phenomenon. Recent studies into automatic activation of memory traces for spoken words of different lexical frequencies suggest that less frequently used items possess less integrated memory traces and therefore take longer to activate; this activation lag manifests itself as a delayed peak latency of corresponding ERP responses (Aleksandrov et al., 2011; Shtyrov et al., 2011). The current findings are in line with this: as the novel word forms are certainly not a frequently used item in the subjects' lexicon, intrinsic neural connections in their newly formed memory circuits cannot be as strong as those for the previously known words, which may be a reason for the lag in activation.

Finally, it appears that the pseudo-word activation in size overtakes that elicited by words in the end of the recording session. Although this effect does not reach significance, it may be an additional sign of the ongoing learning process: novel auditory stimuli early in the process of learning have been shown to produce a larger-scale activation, whilst at later stages tuning of neural representations takes place which optimizes the use of neural resources and prunes unnecessarily activation (Kujala et al., 2003).

Here, we used a passive non-attend paradigm approach which has been repeatedly shown to be a sensitive tool for recording lexical memory trace activations (Shtyrov and Pulvermüller, 2007), which also seems to be the case in the current study. Although the lack of attention to stimuli may be suggestive of certain automaticity in the learning process, this issue was not specifically under investigation here and remains to be explored in future studies which could achieve this by systematically modulating attention on stimuli and manipulating stimulus-related tasks.

CONCLUSION

We have recorded event-related potentials elicited in the brain by novel spoken word forms as they are being learnt through passive auditory exposure. We observed a dramatic change in the brain response dynamics within the short exposure session: as the subjects become familarized with the novel word forms, the early (∼100 ms) fronto-central activity they elicit increases in magnitude and becomes similar to that of previously known real words. Acoustically similar real words used as control stimuli show a stable response throughout the recording session, a sign of robustness of existing linguistic representations. Acoustically matched novel non-speech stimuli do not demonstrate a learning-related response increase, suggesting neural specificity of the rapid learning phenomenon to language. These results suggest that the human brain may efficiently form new cortical circuits online, as it gets exposed to novel linguistic patterns in the sensory input. Left-lateralized perisylvian neocortical networks appear to be underlying such fast mapping of novel word forms unto the brain's mental lexicon.

ACKNOWLEDGMENTS

Yury Shtyrov is supported by the Medical Research Council (MRC), UK (U.1055.04.014.00001.01, MC_US_A060_0043). The study was supported by the MRC, University of Helsinki, and Academy of Finland. The author wishes to thank Teija Kujala, Pasi Piiparinen, and Friedemann Pulvermüller for their help at different stages of this study.

REFERENCES

Aleksandrov, A. A., Boricheva, D., Pulvermüller, F., and Shtyrov, Y. (2011). Strength of word-specific neural memory traces assessed electrophysiologically. *PLoS ONE* 6, e22999. doi:10.1371/journal.pone.0022999

Born, J., Rasch, B., and Gais, S. (2006). Sleep to remember. *Neuroscientist* 12, 410–424.

Borovsky, A., Kutas, M., and Elman, J. (2010). Learning to use words: event-related potentials index single-shot contextual word learning. *Cognition* 116, 289–296.

Breitenstein, C., Jansen, A., Deppe, M., Foerster, A. F., Sommer, J., Wolbers, T., and Knecht, S. (2005). Hippocampus activity differentiates good from poor learners of a novel lexicon. *Neuroimage* 25, 958–968.

Carey, S., and Bartlett, E. (1978). Acquiring a single new word. *Papers Rep. Child Lang. Dev.* 15, 17–29.

Catani, M., Jones, D. K., and ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Ann. Neurol.* 57, 8–16.

Corballis, M. C. (2009). The evolution of language. *Ann. N. Y. Acad. Sci.* 1156, 19–43.

Davis, M. H., and Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 3773–3800.

De Diego Balaguer, R., Toro, J. M., Rodriguez-Fornells, A., and Bachoud-Lévi, A.-C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS ONE* 2, e1175. doi:10.1371/journal.pone.0001175

Dollaghan, C. (1985). Child meets word: "fast mapping" in preschool children. *J. Speech Hear. Res.* 28, 449–454.

Friederici, A. (2002). Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci. (Regul. Ed.)* 6, 78–84.

Garagnani, M., Shtyrov, Y., and Pulvermüller, F. (2009). Effects of attention on what is known and what is not: MEG evidence for functionally discrete memory circuits. *Front. Hum. Neurosci.* 3:10. doi:10.3389/neuro.09.010.2009

Gaskell, M. G., and Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition* 89, 105–132.

Ilmoniemi, R. J. (1993). Models of source currents in the brain. *Brain Topogr.* 5, 331–336.

Kaminski, J., Call, J., and Fischer, J. (2004). Word learning in a domestic dog: evidence for "fast mapping." *Science* 304, 1682–1683.

Kujala, A., Huotilainen, M., Uther, M., Shtyrov, Y., Monto, S., Ilmoniemi, R. J., and Näätänen, R. (2003). Plastic cortical changes induced by learning to communicate with non-speech sounds. *Neuroreport* 14, 1683–1687.

Majerus, S., Van der Linden, M., Collette, F., Laureys, S., Poncelet, M., Degueldre, C., Delfiore, G., Luxen, A., and Salmon, E. (2005). Modulation of brain activity during phonological familiarization. *Brain Lang.* 92, 320–331.

Markson, L., and Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature* 385, 813–815.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457.

Mestres-Misse, A., Rodriguez-Fornells, A., and Munte, T. F. (2007). Watching the brain during meaning acquisition. *Cereb. Cortex* 17, 1858–1866.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113.

Paulesu, E., Vallar, G., Berlingeri, M., Signorini, M., Vitali, P., Burani, C., Perani, D., and Fazio, F. (2009). Supercalifragilisticexpialidocious: how the brain learns words never heard before. *Neuroimage* 45, 1368–1377.

Pittman, A. L. (2008). Short-term word-learning rate in children with normal hearing and children with hearing loss in limited and extended high-frequency bandwidths. *J. Speech Lang. Hear. Res.* 51, 785–797.

Pulvermüller, F., and Shtyrov, Y. (2006). Language outside the focus of attention: the mismatch negativity as a tool for studying higher cognitive processes. *Prog. Neurobiol.* 79, 49–71.

Pulvermüller, F., and Shtyrov, Y. (2009). Spatiotemporal signatures of large-scale synfire chains for speech processing as revealed by MEG. *Cereb. Cortex* 19, 79–88.

Pulvermüller, F., Shtyrov, Y., and Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain Lang.* 110, 81–94.

Rauschecker, A. M., Pringle, A., and Watkins, K. E. (2008). Changes in neural activity associated with learning to articulate novel auditory pseudowords by covert repetition. *Hum. Brain Mapp.* 29, 1231–1242.

Rauschecker, J. P., and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724.

Rosburg, T., Trautner, P., Boutros, N. N., Korzyukov, O. A., Schaller, C., Elger, C. E., and Kurthen, M. (2006). Habituation of auditory evoked potentials in intracranial and extracranial recordings. *Psychophysiology* 43, 137–144.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., and Weiller, C. (2008). Ventral and dorsal pathways for language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18035–18040.

Sharon, T., Moscovitch, M., and Gilboa, A. (2011). Rapid neocortical acquisition of long-term arbitrary associations independent of the hippocampus. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1146–1151.

Shtyrov, Y. (2010). Automaticity and attentional control in spoken language processing: neurophysiological evidence. *Ment. Lex.* 5, 255–276.

Shtyrov, Y., Kimppa, L., Pulvermuller, F., and Kujala, T. (2011). Event-related potentials reflecting the frequency of unattended spoken words: a neuronal index of connection strength in lexical memory circuits? *Neuroimage* 55, 658–668.

Shtyrov, Y., Kujala, T., and Pulvermüller, F. (2010a). Interactions between language and attention systems: early automatic lexical processing? *J. Cogn. Neurosci.* 22, 1465–1478.

Shtyrov, Y., Nikulin, V. V., and Pulvermüller, F. (2010b). Rapid cortical plasticity underlying novel word learning. *J. Neurosci.* 30, 16864–16867.

Shtyrov, Y., Pihko, E., and Pulvermüller, F. (2005). Determinants of dominance: is language laterality explained by physical or linguistic features of speech? *Neuroimage* 27, 37–47.

Shtyrov, Y., and Pulvermüller, F. (2007). Language in the mismatch negativity design: motivations, benefits and prospects. *J. Psychophysiol.* 21, 176–187.

Valo, M. (1994). *Käsitykset ja vaikutelmat äänestä. Kuuntelijoiden arviointia radiopuheen äänellisistä ominaisuuksista.* Jyväskylä: Jyväskylän yliopisto.

Virtanen, J., Rinne, T., Ilmoniemi, R. J., and Näätänen, R. (1996). MEG-compatible multichannel EEG electrode array. *Electroencephalogr. Clin. Neurophysiol.* 99, 568–570.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 August 2011; paper pending published: 26 August 2011; accepted: 01 November 2011; published online: 21 November 2011.

Citation: Shtyrov Y (2011) Fast mapping of novel word forms traced neurophysiologically. Front. Psychology ***2****:340. doi: 10.3389/fpsyg.2011.00340*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 21 May 2012
doi: 10.3389/fpsyg.2012.00158

The benefits of executive control training and the implications for language processing

Erika K. Hussey [1,2] **** and Jared M. Novick*** [1,2]

[1] *Program in Neuroscience and Cognitive Science, Department of Psychology, University of Maryland, College Park, MD, USA*
[2] *Center for Advanced Study of Language, University of Maryland, College Park, MD, USA*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Silvia P. Gennari, University of York, UK
Marina Bedny, Massachusetts Institute of Technology, USA

****Correspondence:***
Erika K. Hussey, Program in Neuroscience and Cognitive Science, Department of Psychology, University of Maryland, 1147 Biology-Psychology Building, College Park, MD 20916, USA.
e-mail: ehussey@umd.edu

Recent psycholinguistics research suggests that the executive function (EF) skill known as conflict resolution – the ability to adjust behavior in the service of resolving among incompatible representations – is important for several language processing tasks such as lexical and syntactic ambiguity resolution, verbal fluency, and common-ground assessment. Here, we discuss work showing that various EF skills can be enhanced through consistent practice with working-memory tasks that tap these EFs, and, moreover, that improvements on the training tasks transfer across domains to novel tasks that may rely on shared underlying EFs. These findings have implications for language processing and could launch new research exploring if EF training, within a "process-specific" framework, could be used as a remediation tool for improving general language use. Indeed, work in our lab demonstrates that EF training that increases conflict-resolution processes has selective benefits on an untrained sentence-processing task requiring syntactic ambiguity resolution, which relies on shared conflict-resolution functions. Given claims that conflict-resolution abilities contribute to a range of linguistic skills, EF training targeting this process could theoretically yield wider performance gains beyond garden-path recovery. We offer some hypotheses on the potential benefits of EF training as a component of interventions to mitigate general difficulties in language processing. However, there are caveats to consider as well, which we also address.

Keywords: cognitive training, executive function, conflict resolution, process-specificity, language processing, ambiguity resolution

INTRODUCTION

Cognitive control, also called executive function (EF), refers to a cluster of mental processes that permit the flexible adjustment of thoughts and actions across domains, allowing individuals to adapt to new rules and guide the selection of task-relevant over task-irrelevant information in an environment that varies continuously (Miller and Cohen, 2001). As we navigate our surroundings, we can frequently rely on a set of highly regularized functions that render certain tasks like driving a car or skimming a magazine article relatively automatic. Sometimes, however, new instructions or conflicting information compels us to override these reflexive actions and instead consider what might otherwise be a disfavored (or atypical) response. For instance, a resident of Chicago may be in the habit of making a legal right turn on red when driving at home, but this routine behavior could result in a costly ticket when she visits New York City, where turning on red is strictly prohibited! Likewise, imagine reading the following sentence upon skimming a magazine: *at the restaurant, the interns discussed the bill before suggesting edits to the senator.* One might initially interpret the word "bill" to mean the list of charges incurred for the meal, rather than its intended (though less common) interpretation, namely a draft piece of legislation. On the surface, both examples are quite different, but conceivably induce a similar experience: the detection of an incompatibility and the ensuing need to rein-in a highly familiar, yet currently inappropriate cognitive reaction (e.g., refrain from turning; revise the more frequent meaning, but current misanalysis, of "bill"). Such "conflict resolution" functions are an essential part of cognitive control (Botvinick et al., 2001) and help adapt information-processing strategies so individuals can regulate behavior in view of ever-changing goals, new contexts, or situation-specific demands.

As many researchers have argued, EFs encompass a *collection* of cognitive processes that help guide goal-directed behavior; that is, cognitive control is not a unitary construct but comprises separable components (Norman and Shallice, 1986; Botvinick et al., 2001; Miller and Cohen, 2001). In addition to the conflict-resolution processes outlined above, other EFs include task-switching, updating, and information monitoring, each of which can operate over visual, spatial, or verbal domains (Smith and Jonides, 1999; Miyake et al., 2000; Friedman and Miyake, 2004) and thus may be recruited across a variety of tasks including selective attention, decision-making, working memory (WM), error monitoring, and language processing (Botvinick et al., 2001; Thompson-Schill et al., 2005; Badre and Wagner, 2007; inter alia). With regard to conflict-resolution functions in particular, converging data from neuropsychological patients and brain-imaging studies of healthy adults suggest that, across a range of WM, attention, and language tasks, posterior regions of left ventrolateral prefrontal cortex (VLPFC) commonly support the ability to resolve among competing sources of evidence, regardless of domain (Thompson-Schill et al., 2005).

In this paper, we discuss how a burgeoning literature demonstrates that EFs can be trained through ample practice – that such abilities are seemingly not fixed, but malleable – and that performance increases throughout the course of training generalize to novel tasks that were not part of the training protocol. Some examples of transfer include benefits on unpracticed tasks tapping fluid intelligence (Jaeggi et al., 2008), working-memory updating (Dahlin et al., 2008; Li et al., 2008), and task-switching (Karbach and Kray, 2009) – that is to say, transfer benefits have been observed across a *range* of EF.

We are especially interested in the implications that these training-transfer findings have for language processing under conditions of conflict, given that domain-general conflict-resolution and cognitive-control functions have been associated with assorted linguistic abilities including the resolution of lexical (Bilenko et al., 2009; Copland et al., 2009; Khanna and Boland, 2010) and syntactic ambiguities (Novick et al., 2005), verbal fluency (Robinson et al., 1998; Kan and Thompson-Schill, 2004; Novick et al., 2009; Schnur et al., 2009), and perspective-taking during natural dialog (Brown-Schmidt, 2009; Nilsen and Graham, 2009; for reviews, see Novick et al., 2005; Novick et al., 2010). Thus, in the *hypothesis* section, which details the potential implications of EF training and transfer effects on language use, we consider a *theory* based on evidence that left VLPFC-supported conflict resolution is the kind of cognitive-control function of principal relevance to these particular linguistic tasks (see e.g., Novick et al., 2005). We couch our hypotheses within a *process-specific* account (see Dahlin et al., 2008; Shipstead et al., 2010, 2012), which in the training literature posits that post-intervention, performance increases on novel tasks largely depends on the extent of overlap between the training and transfer measures, both in terms of the shared cognitive processes and underlying neural systems needed to complete them. That is, if a certain component of EF (e.g., conflict resolution) is targeted and improved through training, then transfer measures relying on common processes should be influenced accordingly, irrespective of domain. In view of this, we will focus our discussion on a few language comprehension and production tasks that fit within the VLPFC-mediated process-specific function typically referred to as "conflict resolution." However, in the discussion, we acknowledge other brain systems involved in a wider array of EFs, and consider briefly the implications for training and the effects on language.

As sketched in the driving and reading examples earlier, when we talk about conflict (or interference), we are referring to conditions that contain the presence of mismatched or incongruent sources of evidence. Specifically, "conflict" designates cases in which current situation-specific demands generate an incompatibility between how an input stimulus should be characterized (dubbed *representational conflict*), given how the input is normally considered. Such conflict is often called "prepotent conflict," because individuals must override their dominant (prepotent) biases in support of atypical alternatives (Botvinick et al., 2001). For instance, the Stroop task is a canonical representational conflict task involving the need to countermand a prepotent bias that is generated by a lexical representation (which gives rise to an automatic reading response), in favor of a perceptual (color) representation. A comparable type of representational conflict occurs in the form of "underdetermined conflict," in which multiple candidate representations are equally reasonable and thus compete for selection (Botvinick et al., 2001). Importantly, brain-imaging findings suggest separable neuroanatomical involvement for representational conflict versus *response conflict* (or *response selection*; see Milham et al., 2001; Nelson et al., 2003). Our major focus here is on the implications of conflict-resolution training at the representational level on particular language-performance measures such as lexical and syntactic ambiguity resolution (in comprehension) and verbal fluency (in production). Both prepotent and underdetermined representational conflicts recruit posterior regions of VLPFC (Brodmann areas 44 and 45) across language and memory domains, meeting the requirements for a test of process-specificity (see Novick et al., 2010; see also Milham et al., 2001 and Nelson et al., 2003, which demonstrate VLPFC recruitment for representational conflict resolution but anterior cingulate recruitment for response-level conflict resolution).

Generally, we believe that – considering the mounting evidence showing the effectiveness of various types of EF training in different populations (Klingberg et al., 2005; Westerberg et al., 2007; Jaeggi et al., 2011) – there is room to establish new research investigating if EF training protocols that focus on selective subprocesses (i.e., representational conflict resolution) could be used successfully as an intervention technique to mitigate problems in general language use that arise under high-EF (i.e., high-conflict) demands.

Indeed, there is tantalizing evidence supporting process-specific transfer to conflict-related language measures, drawn not from a long-term training paradigm *per se*, but rather from another type of intervention designed to *fatigue* selective cognitive processes common to WM and language processing tasks. These so-called "resource depletion models" offer an interesting framework to understand *negative* transfer to tasks relying on temporarily exhausted EFs shared across ostensibly different domains (Van der Linden et al., 2003; Persson et al., 2007). That is, rather than boosting general-purpose EFs through long-term practice, as is the case with training studies, resource depletion paradigms rely on short-term "overuse" of a particular cognitive process. For example, after performing a complex task that places high demands on EF capacities, these resources are rendered temporarily unavailable for continued use; therefore, performance *decreases* on transfer measures that rely on the common "worn out" EF (Van der Linden et al., 2003; Persson et al., 2007; see also Snyder et al., 2010 for similar findings among anxious individuals).

In one study (Persson et al., 2007), conflict-resolution abilities were fatigued through an intensive session of an item-recognition task with high conflict-resolution demands. In this task, participants indicated whether a probe item (e.g., *C*) appeared in an immediately prior memory set (e.g., *r, f, c, l*; see Monsell, 1978). Frequently, subjects could respond correctly due to familiarity alone: familiar probes required a "yes" response and unfamiliar ones a "no" response. However, relying on familiarity on some "no" trials was prone to error, because they contained a probe (e.g., *G*) that was not among the current memory set (*j, p, v, m*) but *was* among the items in the *prior* trial (*g, k, v, p*). Thus, these trials required subjects to override a prepotent familiarity bias (and "yes" response) and instead re-characterize the probe stimulus as

"familiar-but-irrelevant," and respond "no." Such "recent-no" trial types, when compared to "non-recent-no" trials (when the probe did not appear in either the current or preceding sets) routinely recruit left posterior VLPFC (Jonides and Nee, 2006). Important for the current discussion, after subjects completed this task and "fatigued" the conflict-resolution process, they subsequently demonstrated selective performance decline on VLPFC-mediated, high-conflict conditions on a verbal fluency task, in which they had to generate an associated verb to a given noun (e.g., *scissors* → *cut*; high-conflict items had many possible associated verbs, like *ball* → *kick*, *throw*, *catch*, *bounce*, and thus contained underdetermined response conflict; see Thompson-Schill et al., 1998). This pattern of negative transfer was not observed for (1) subjects who received exposure to only low-conflict trials during their intensive practice session (i.e., no recent-no trials were present); or (2) individuals who practiced a different task before the verb generation task, namely a stop-signal task that recruits mainly right-hemisphere networks and a different subcomponent of EF (response inhibition; see also Friedman and Miyake, 2004). Together, this suggests that the process-specificity observed across intervention and transfer tasks operates on a short time scale, such that as conflict resolution is temporarily depleted, other tasks relying on shared cognitive and neural resources are affected accordingly.

Although these effects are transient, the selective transfer findings are nonetheless critical: they demonstrate that conflict resolution abilities are at least temporarily malleable, and this malleability can subsequently affect language processing under similar conditions of high conflict. Consequently, we ask: considering evidence for process-specific transfer, on a short time scale, across memory and language tasks that commonly rely on VLPFC-mediated conflict-resolution functions, might one observe longer-term effects on language measures as well, when conflict resolution is boosted via extensive practice? That is, can we observe *positive* transfer – namely, performance *increases* – when individuals consistently train conflict-resolution functions over time? We hypothesize that the answer should be yes, given the evidence that other EFs (e.g., task-switching, etc.) are both trainable and transferrable. Indeed, work from our lab demonstrates reliable transfer to syntactic ambiguity resolution in healthy adults, where individuals who have undergone extensive conflict-resolution training fare significantly better at revising early misinterpretations than their untrained counterparts (Hussey et al., 2010; Novick et al., submitted for publication). Additionally, on the basis of the theory that posterior regions of VLPFC support conflict resolution across domains, such displays of transfer, we hypothesize, might clearly extend beyond just "garden-path" recovery, given the putative role of conflict-resolution in several other measures of language processing.

Although we outline below some potential benefits of conflict-resolution training on language use, we also discuss some caveats that should be considered, including individual differences in training success (not everyone responds to training or achieves similarly high levels, cf. Chein and Morrison, 2010; Jaeggi et al., 2011), limitations that may be involved in training special populations, and the need for explicit linking hypotheses between training and any expected transfer: namely, there must be a theory that bridges the hypothesized underlying cognitive processes from one task to another (i.e., from an intervention task to a transfer task). Transfer from training to untrained assessment tasks cannot be expected, or explained, without a well-formulated process-specific theory (Shipstead et al., 2010, 2012). To this end, we also speculate that the magnitude of transfer effects is contingent upon the *degree* to which a targeted EF contributes to and shares critical features with an outcome measure. This is particularly important if, as some researchers suggest, EF is not a unitary construct but is comprised of separable, multi-component processes such as conflict resolution, updating, and task-switching (Miyake et al., 2000; Persson et al., 2007; Dahlin et al., 2008).

As outlined in this hypothesis and theory piece, we integrate the extant training and psycholinguistic literatures to develop testable hypotheses from an emerging picture within the EF training research. The following section begins with a brief review of cognitive training studies demonstrating transfer to novel tasks that are ostensibly different from those practiced during the training regimens, but share specific processing demands. We then turn to research on the role of conflict resolution in language use, sketching some hypotheses and implications the training findings have for new work aimed at improving language processing under high-EF – particularly high-conflict-resolution – demands. That is, if conflict-resolution is malleable (which seems to be the case given the resource depletion work outlined above), we *hypothesize* that training such processes should also show transfer to untrained measures of conflict resolution within the linguistic domain, patterning with other training-transfer findings. The *theory* bolstering this claim comes from work (drawn from patients, children, and brain-imaging studies of adults) indicating that conflict-resolution and cognitive-control measures play an important role in language tasks that we outline below.

EXECUTIVE FUNCTION TRAINING AND ITS TRANSFER ACROSS COGNITIVE DOMAINS: A BRIEF REVIEW

A recent flurry of research is devoted to testing if general-purpose cognitive abilities can be enhanced through consistent practice with WM tasks that recruit brain regions within the cortico-striatal network key to executive functioning. Although interventions geared toward improving psychological faculties, specifically intelligence, were pioneered decades ago (see Feuerstein, 1980), Klingberg and colleagues have recently reinstated the notion by training domain-general cognitive abilities as a means to remediate populations with diminished WM resources including stroke patients (Westerberg et al., 2007), children with attention deficit hyperactivity disorder (Klingberg et al., 2005), and older adults (Brehmer et al., 2011). Ever since, cognitive training programs have undergone significant study, particularly in healthy adults, to examine whether normally functioning individuals' EF abilities can be improved, and what generalized outcomes consistent training might have on everyday performance on non-trained tasks. To this end, researchers have been investigating questions related to dosage-dependence (does more practice yield more transfer?; Jaeggi et al., 2008), the extent to which training transfers to untrained but related measures (Li et al., 2008; Karbach and Kray, 2009; Chein and Morrison, 2010; Morrison and Chein, 2011), if training tasks must adapt to individuals' performance

to be effective (Klingberg et al., 2005; Brehmer et al., 2011), and individual differences in training success (Jaeggi et al., 2011).

Here, we focus on the extent to which training generalizes to novel tasks. The typical training study is designed as a pre/post longitudinal experiment in which subjects are assessed on some cognitive capacity immediately before and again after an extensive intervention. In some cases, the intervention comprises practice with a single training task (Dahlin et al., 2008; Jaeggi et al., 2008, 2011; Li et al., 2008), whereas in others, a battery of training tasks is administered (Klingberg et al., 2002, 2005; Karbach and Kray, 2009). Regardless, the training tasks are different from those completed at the pre/post assessment sessions, with the intervention component typically lasting for several hours distributed over a few weeks. Upon conclusion of the regimen, trainees return to the lab and complete follow-up assessments, namely complementary versions of the tasks that were done just prior to training, to evaluate whether performance on assessments has reliably improved, thereby providing evidence for "transfer."

Transfer has been documented for untrained tasks that share obvious features with well-practiced training tasks, an effect sometimes referred to as "near-transfer." For instance, performance increases on WM training tasks generalize to structurally similar (but new) WM assessments (Li et al., 2008; Karbach and Kray, 2009; see below). However, "far transfer" can also be observed, namely to assessments that appear, on the surface, to be wildly different from the training tasks completed throughout the intervention regimen (Kloo and Perner, 2003; Dahlin et al., 2008; Jaeggi et al., 2008, 2011). This latter form of transfer is possible provided that training and assessment tasks share certain essential underlying EFs (as well as overlapping neural resources; see Jonides, 2004; Shipstead et al., 2010, 2012).

NEAR-TRANSFER OF TRAINING

Near-transfer effects emerge when the nature of the processed information – including stimulus type, task structure, and response type – is similar across training and assessment tasks (but see Morrison and Chein, 2011 for an alternative definition of near-transfer). For instance, in one report (Li et al., 2008), trainees practiced a spatial 2-back task, during which they had to monitor the locations of sequentially presented squares on a 3×3 grid and respond whenever the current location matched the location seen two trials earlier. Compared to a no-contact control group, trained participants demonstrated post-intervention improvements on a spatial 3-back task, providing evidence for near-transfer to a more difficult, but otherwise identical task. Another type of near-transfer occurs when the *type* of information (i.e., the stimuli) being processed is changed across training and transfer tasks, while the response-level requirements remain constant, resulting in a structural continuity between both tasks. For example, in the same study by Li et al. (2008), trainees also improved on *numeric* 2- and 3-back tasks, where instead of remembering locations on a grid, subjects indicated when a serially presented number (0–9) matched the identity of a number presented two (or three) trials previously. The authors argued that transfer to a numeric *n*-back task provided support for a task-specific response strategy shared across stimulus modalities: Although the spatial 3-back and numeric *n*-back tasks differ from the spatial 2-back training task, all require the same basic strategy, namely, information must be monitored and updated in a predictable fashion.

In addition to the above findings, Karbach and Kray (2009) observed that increases in task-switching abilities – an EF based on mental shifting across different goals or rules – as a consequence of training generalizes to performance on novel tasks with similar switching demands. Specifically, their training regimen involved making two-alternative forced-choice judgments about pictures (trees/flowers), based on two separate characteristics (e.g., identity vs. color), such that the relevant characteristic (or rule) changed predictably across trials. Stimulus types (fish/birds, trees/flowers, sports/music, planes/cars) and response categories (identity, number, color, and rotation) varied across sessions within the training regimen. An assessment of near-transfer involved responding to a novel set of stimuli (fruits/vegetables) using number and identity as response categories; compared to a non-switching active-control group, the task-switching trainees showed greater posttest improvement in switching costs, i.e., the difference in response time on switch (color followed by identity judgment) vs. non-switch trials.

These examples highlight two sources of near-transfer: training and outcome measures tap the same underlying EFs (e.g., monitoring and updating), and both tasks provoke similar processing demands through a shared task structure (task-specific aspects). Consequently, it is difficult to disentangle the source of near-transfer effects, as two possibilities may account for any observed pre/post changes: (1) the trained EF shared by both tasks may have been improved, or (2) a task-specific strategy may have been developed. Indeed, in cases of near-transfer, the training and transfer tasks need not tap the same underlying EFs, since transfer could occur simply with improvements at task-specific aspects of the paradigm. Near-transfer effects might be unsurprising: practicing an *n*-back task improves *n*-back performance, and therefore transfers to other *n*-back tasks (perhaps regardless of domain); likewise, practicing a categorization task-switching task generalizes to a similar task with novel categories. But, the extent to which these near-transfer effects are driven by the shared EFs across training and assessment tasks, the surface-level features (stimulus or response characteristics) that are isomorphic between both sets of tasks, or through a combination of both factors is unknown.

FAR-TRANSFER OF TRAINING

Training studies designed to show far-transfer effects help to elucidate the role of shared EFs; by design, the surface-level properties – stimuli or required responses – of the training and assessment tasks are quite different. Consequently, contrary to near-transfer findings, far-transfer effects are assumed *not* to rely heavily on the structural (task-specific) similarities across training and assessment tasks, and instead result mostly from improvements on underlying EFs important to both the training and assessment measures (Shipstead et al., 2010). In other words, the goal of far-transfer training is rooted in improvement of specific processes engaged during tasks with dissimilar structures, often spanning domains (again, sometimes referred to as *process-specific training*).

For instance, in one set of studies, subjects practiced a dual *n*-back memory task involving simultaneous updating of shape locations and the identity of heard letters, such that a target was

defined as an item repeating *n*-trials previously in either modality (Jaeggi et al., 2008). Trainees showed subsequent improvements on Raven's Advanced Progressive Matrices, a transfer task that requires participants to select a textured shape from a set of possible response items, which fits a sequence of other textured shapes to complete a particular pattern with one absent piece (Jaeggi et al., 2008, 2011). The response and surface-level properties of *n*-back and Raven's are distinct, as one task involves monitoring a continuous stream of letters or block locations for familiar instances, and the other requires reasoning to identify the missing element that completes a 4×4 matrix containing orderly patterns across rows and columns; thus, to observe transfer, there must be an underlying process common to both tasks that is enhanced through intensive *n*-back training. The authors reasoned that this shared process centered around a common need to employ attentional control, such that their training procedure – which forced trainees to practice constant shifting of attention to new stimuli – facilitated this ability, thereby enabling transfer to Raven's, which similarly involves updating and selection among multiple representations (via the control of attention). Importantly, because the training and transfer measures were characteristically so different, the authors argued that task-specific elements could not explain the observed generalization, effectively ruling out near-transfer as an explanation for their findings. Rather, training boosted a part of the EF system – here, multiple-task management and attentional control processes – important for a range of cognitive tasks, including Raven's performance. Indeed, separate work demonstrates that *n*-back and Raven's activate a similar network of neural regions, providing additional support for resources common to both tasks (Burgess et al., 2011).

Additional evidence of process-specific training comes from demonstrations of selective far-transfer from an updating task (letter running-span) to a structurally different assessment measure (number *n*-back) that requires a similar updating EF; critically though, such transfer was not demonstrated on the Stroop task, which relies on a separable EF – conflict resolution (Dahlin et al., 2008). During the letter running-span task, participants must recall the last four items of a study list that terminates unexpectedly, forcing them to continuously update the correct response set from a fleeting memory store; similarly, their version of *n*-back required subjects to monitor and refresh representations as new information is processed and deemed relevant. Running-span and a standard number *n*-back task recruit similar striatal regions, corroborating their underlying reliance on a common EF. Contrastingly, tasks requiring conflict resolution, like Stroop, require subjects to re-characterize an automatized response (reading) in order to promote atypical, but task-relevant information (color name); such tasks rely on a separable neural profile (compared to that required for updating tasks) including a network of frontal and parietal regions. Dahlin et al. (2008) demonstrated that training on running-span confers benefits to assessment measures that share updating demands and corresponding neurological profiles (*n*-back), while those with little or no such overlap (Stroop) show negligible improvement. In sum, the amount of far-transfer to untrained tasks following intervention depends on the degree of overlap among cognitive and neural resources shared by the training and the transfer tasks.

Given these training and far-transfer effects for a range of EFs (e.g., attention control, memory updating), one might also hypothesize that transfer from general-purpose EF training to certain tasks of language processing might occur as well. That is, the language tasks are not trained *per se*, but tap particular cognitive functions (conflict resolution) that may be trainable through an extensive regimen targeting common processes (or neural resources). As hypothesized below, the result could be an alleviation of language processing difficulty under conditions that place heavy demands on the EF system in healthy, and perhaps even in special populations. We focus on a select few of these language conditions in the following section, concentrating specifically on a functional-anatomical association between conflict-resolution processes of EF, and regions within left VLPFC that support them (for an extensive review, see Novick et al., 2010). We sketch how this association is important for production and comprehension abilities in healthy adults, young children, and patients with circumscribed VLPFC damage.

THE ROLE OF EXECUTIVE FUNCTION IN LANGUAGE USE: HYPOTHESES AND IMPLICATIONS FOR TRAINING

One priority in psycholinguistics has been to study how non-linguistic cognitive abilities contribute to language production and comprehension. EF abilities have emerged as a candidate characteristic, defining in part those individuals who can better coordinate rapidly among multiple sources of linguistic (syntactic, semantic) and extra-linguistic (pragmatic, contextual) evidence across a range of communicative tasks. Given the breadth of work on various EFs for language, we focus only on the role of conflict-resolution training for a handful of language tasks. As sketched in the introduction, conflict resolution refers to the re-characterization of information in the face of competing sources of evidence. Regarding language processing, good conflict-resolution skills enable readers and listeners to avoid comprehension errors in the face of ambiguity (e.g., by consulting top-down evidence to override misinterpretations), produce the right word among competing options, and take an interlocutor's perspective when assessing common-ground information during natural, unscripted dialog (see Novick et al., 2005, 2010). Indeed, patients with circumscribed damage to left posterior VLPFC consistently underperform on high-conflict conditions on non-linguistic tasks such as Stroop and the "recent-no" task described above (Hamilton and Martin, 2005). Moreover, this general conflict-resolution disorder in patients has been tied to their concomitant deficits on language tasks that generate similar conflict-resolution demands, for example, when dominant meanings of lexical ambiguities must be countermanded (Bedny et al., 2007), when initial interpretations of syntactic ambiguities must be reprocessed (Novick et al., 2005, 2009), or when object names must be selected among categorical competitors (Schnur et al., 2009). As such, by training general-purpose conflict-resolution abilities – supported by regions within VLPFC – in healthy adults, we hypothesize that there should be systematic improvements in high-conflict conditions on language tasks requiring shared demands for conflict resolution. Below, we provide examples of when conflict-resolution abilities appear to

interact with particular language processing skills and outline the implications these associations have for process-specific training.

SYNTACTIC AMBIGUITY RESOLUTION

Theory

Readers and listeners process sentences in real-time, committing to an interpretation incrementally as words and phrases are encountered moment-by-moment (Altmann and Kamide, 1999; Tanenhaus, 2007). One consequence of incremental processing is temporary ambiguity: the first analysis individuals assign sometimes turns out wrong. Cognitive control has been tied to individuals' ability to adjust interpretations when late-arriving evidence signals that their initial analysis was incorrect (Novick et al., 2005). Such cases of conflict (the so-called "garden-path effect") elicit temporary processing difficulty in reading (Frazier and Rayner, 1982; Staub and Rayner, 2007; *inter alia*) and confusion during spoken comprehension (Tanenhaus et al., 1995). Individuals must then engage in a process that permits them to revise and capture the intended interpretation.

Evidence for the role of conflict-resolution in this recovery process comes from populations with underdeveloped or impaired cognitive control such as young children (whose PFC development is protracted; see Huttenlocher and Dabholkar, 1997) and patients with focal damage to left posterior VLPFC. Both populations fail to initiate cognitive-control functions across assorted non-syntactic measures (e.g., Stroop, the recent-no, and other analogous tasks; e.g., Hamilton and Martin, 2005; Khanna and Boland, 2010), and both groups similarly fail to revise sentence interpretations following early misanalysis (Trueswell et al., 1999; Weighall, 2008; Novick et al., 2009; see also Christianson et al., 2006 for similar patterns in older adults). The linking assumption is that the discovery of a misinterpretation deploys conflict-resolution to resolve the incompatibility between representations of sentence meaning: the one initially assigned and the one in need of recovery, similar to the controlled processes required to resolve conflict during incongruent Stroop trials, or interference from familiar but currently irrelevant items in the "recent-no" task (Hamilton and Martin, 2005; Novick et al., 2005, 2010). Interestingly, healthy adults undergoing functional neuroimaging demonstrate co-localized neural activity within left posterior VLPFC when performing both syntactic and non-syntactic tasks requiring conflict resolution, corroborating the necessary involvement of shared, domain-general processes presumed from special populations (January et al., 2009; Ye and Zhou, 2009).

Hypothesis

This convergence of findings suggests an opportunity to alleviate the processing difficulty associated with temporary ambiguities that arise during sentence processing by targeting the EFs (through training) that appear to be domain-general, i.e., common across certain syntactic and non-syntactic tasks. We tested this hypothesis in a study in which healthy trainees completed pre/post reading assessments involving syntactically ambiguous sentences susceptible to misanalysis (Hussey et al., 2010; Novick et al., submitted for publication). We hypothesized that practicing a performance-adaptive *non-linguistic* task requiring conflict-resolution processes – the *n*-back memory task with lures (see below) – would endow trainees with improved abilities essential to re-interpreting garden-path sentences. (Performance adaptation means that as subjects reached a certain criterion, task difficulty increased dynamically in terms of *n* and the number of lures present.) Similar to the processing demands of the recent-no task, our training task required participants to re-characterize stimulus representations in real-time. Specifically, subjects completed a version of *n*-back during training that contained interference lures, or items that match in target-identity but appeared in non-*n*-positions. For example, in the sequence *G-P-K-G*, the second *G* would be a target in a 3-back condition because it matches the 3-back stimulus. However, in the sequence *G-P-K-L-G*, the second G would be a "lure" in a 3-back condition because it matches the stimulus presented four, not three, items back (Gray et al., 2003; Burgess et al., 2011). We argued (as have others) that the familiarity of lure items forces participants to engage conflict-resolution functions to override a familiarity bias and the tendency to respond "target" to familiar representations; instead, subjects must re-characterize familiar letters in non-*n* locations as non-targets (thus lures are akin to "recent-no" trials in the item-recognition task). Importantly, neuroimaging work (Gray et al., 2003) demonstrates that lure trials activate VLPFC resources that are also recruited during high-conflict language processing tasks. This finding suggests that practicing an *n*-back task with lures may lead to improvements not just on that task, but also in resolving competing interpretations of syntactically ambiguous sentences.

To examine process-specific training-related changes in sentence processing, readers' eye movements were recorded; we were primarily interested in the effect of training on processing difficulty, particularly in sentence regions that introduced new evidence signaling an incompatibility with individuals' early interpretations (i.e., disambiguating regions that induce conflict). Readers also answered comprehension questions, the responses to which indexed a failure to ultimately override their original misanalysis (Christianson et al., 2006). We found three important patterns: (1) those trainees who responded most to *n*-back practice – reflected in steady performance gains throughout the regimen – demonstrated significantly improved comprehension accuracy at posttest for ambiguous (but not unambiguous) materials, whereas the untrained controls and non-responsive trainees did not; (2) responsive trainees' reading times were reliably faster at posttest, acutely in disambiguating regions of ambiguous sentences, but not in other regions, reflecting less processing difficulty post-training upon encountering conflicting evidence – the control group and non-responders demonstrated no test-retest change; and (3) trainees' performance improvement on *n*-back-with-lures – and no other training task administered as controls – predicted the increases they achieved in garden-path recovery.

The selectivity of these findings is of particular interest, because trainees exhibited improvements only on the language materials where conflict-resolution processes are hypothesized to trigger (unambiguous materials did not involve the need to employ control to revise interpretations, and no test-retest changes in accuracy or reading times were found in this condition). Further, these pre/post improvements were accounted for only by individual training gains on the *n*-back-with-lures task – i.e., a task requiring conflict resolution – and no other well-practiced

WM task completed during intervention (participants also trained on tasks tapping visuo-spatial and verbal WM functions without conflict-resolution demands). Importantly, many researchers argue (D'Esposito and Postle, 1999; Kane and Engle, 2000) that there are some tasks of WM that tap non-mnemonic functions, such as the need to resolve conflicting representations, which is a general-purpose skill necessary for some (not all) WM tasks and some language tasks like syntactic ambiguity resolution (Novick et al., 2005).

Overall, the patterns are consistent with the idea that the ability to recover from misinterpretation can be enhanced by training domain-general EFs common to some tasks of language processing and some tasks of WM. These findings indicate that within the right framework, and having appropriate linking hypotheses, EF training may be a viable way to improve language use under certain conditions through tests of far-transfer. Open questions remain about the trainability of special populations – particularly if training VLPFC patients and young children with poor conflict-resolution skills will result in improved cognitive control, extending to an enhanced ability to recover from parsing misanalyses. But the opportunity to test such ideas is ripe. To our knowledge, this study is the first to investigate the impact of EF training on the processes that commonly contribute to language comprehension. As sketched below, conflict-resolution abilities are associated with various other specific language processing tasks, leaving room to explore the effects of training on language use more generally.

LEXICAL AMBIGUITY RESOLUTION

Theory

Research examining comprehension at the single-word level suggests a role for conflict resolution when the dominant meaning of an ambiguous word (e.g., *bill*, as the tab issued by a restaurant) must be overridden to retrieve its subordinate meaning (an outline of a prospective law; Bedny et al., 2007). Questions posed in this literature examine whether good conflict-resolution skills enable context-dependent meaning selection, and conversely, whether poor abilities impair it. Researchers have found that better conflict resolution is related to young children's contextual sensitivity: context *can* be used by kids to countermand dominant, but inappropriate meanings of an ambiguous word; however, the use of top-down information is largely dependent on the maturity of their EF abilities, as indexed by a separate task of conflict resolution and inhibitory control (Khanna and Boland, 2010). Correspondingly, neuropsychological patients with poor conflict resolution show inadequate lexical ambiguity resolution when the subordinate meaning is activated by local contextual information (Balota and Faust, 2001; Bedny et al., 2007), suggesting that such patients have difficulty suppressing context-inappropriate meanings of ambiguous words (Copland et al., 2009; Vuong and Martin, 2011). Finally, across several studies, regions within VLPFC – the same areas involved in lesion-deficit analyses of patients showing conflict-resolution impairments – are active in healthy adults during lexical-decision tasks necessitating resolution of meaning competition, suggesting that VLPFC-mediated EFs trigger to resolve increased competition associated with accessing the less frequent meaning of an ambiguous word (Bilenko et al., 2009).

Hypothesis

Considering the training results observed for syntactic ambiguity resolution – and therefore assuming that conflict resolution is yet another trainable EF in addition to updating and task-switching – lexical ambiguity resolution abilities may also be enhanced, hypothetically, through conflict-resolution training tasks designed to target EFs central to overriding dominant biases and implementing cognitive control (provided the effects are large enough to observe improvement; this may be particularly true in clinical patients). Future research might test whether EF training, with the right tasks, could garner improvements in integration among top-down contextual and lexical sources of evidence, particularly when these latter sources give rise to multiple conflicting meanings. There are obvious implications for clinical patients with word-comprehension deficits stemming from poor conflict-resolution abilities.

REFERENCE RESOLUTION

Theory

When conversational participants interact, they establish what is known as "common ground," or shared beliefs. Brown-Schmidt (2009) has demonstrated that variations in cognitive-control abilities can explain healthy individuals' occasional inattentiveness to common-ground information; that is, objects visually accessible only to the listener are occasionally (incorrectly) favored as a referential interpretation over objects accessible to both partners. Specifically, individual differences in conflict resolution may determine if a listener can successfully override perspective-inappropriate interpretations of referential ambiguities uttered by their partner. As such, conflict resolution may predict how easily semantic and pragmatic information is integrated in order to rule out incorrect interpretations during natural dialogue.

Indeed, a study testing young children corroborates this account by showing that although 5-year-olds can distinguish common versus privileged knowledge during conversation, the preference for their own perspectives – assessed by gaze duration to inappropriate privileged-ground alternatives – is predicted by measures of conflict resolution and inhibitory control (Stroop, a tapping task, and the *bear/dragon* puppet task), all of which require resolving among conflicting representations by overriding a dominant rule/bias (Nilsen and Graham, 2009). That is, children with poorer cognitive-control demonstrated exaggerated looking times to high-conflict referential alternatives inaccessible to the speaker but hidden (or "privileged") so that only the listener (the child) can see them (e.g., a small duck when "*Look at the duck*" is uttered and competes with the target that is common knowledge, i.e., a large duck). Namely, children with better performance on high-conflict conditions of an inhibitory control task were more likely to override their egocentric view and modify their behavior to be consistent with information shared by both communicative parties, and did so *selectively* for high-conflict items evidenced by spending less time gazing at inappropriate privileged-ground alternatives.

Adults occasionally show similar consideration of perspective-inappropriate interpretations when a speaker utters a referential ambiguity, failing to be sensitive to common-ground information immediately. This behavior is also related to individual

variation in conflict-resolution abilities. For instance, during one "visual-world" task (Brown-Schmidt, 2009), participants assisted the experimenter in revealing the identity of subject-privileged pictures on a display by answering the experimenter's questions. Generally, addressees consulted common-ground information to resolve temporarily ambiguous requests, like, *What's above the horse with the glasses*?, when two horses might be referenced, one wearing glasses and another wearing shoes. If the item above one of the horses (the horse with shoes) was previously grounded, then subjects directed their gaze toward the unmentioned target and the horse (with glasses) located below it, as the ambiguity unfolded. Crucially, however, the degree to which an addressee was able to use perspective information to avoid considering inappropriate interpretations (i.e., understanding the question to mean the already-revealed object) was determined by his Stroop performance. That is, subjects with better conflict-control were quicker to resolve referential conflict by directing their attention away from grounded items and toward previously unmentioned items.

Although conflict-resolution measures account for the individual differences in perspective-taking ability in children and adults, common-ground assessment likely requires multiple different kinds of EF (e.g., memory for perspective). However, it is important to note that the only experimental conditions predicted by Stroop performance are those that impose high conflict-resolution demands.

Hypothesis

This raises the question: if relevant EF skills can be targeted and enhanced via conflict-resolution training (for instance, using a training-appropriate version of the Stroop task as in Brown-Schmidt, 2009), would individuals (particularly children) subsequently be less likely to consider unintended interpretations in cases of referential ambiguity? That is, one might hypothesize that EF training, within a process-specific conflict-resolution framework, will result in a generally sharper ability to promote relevant sources of information like context and pragmatics, and suppress currently irrelevant ones (e.g., one's privileged perspective) through top-down control.

Indeed, there is indirect yet tantalizing support for this. Work by Kloo and Perner (2003) provides evidence for far-transfer across structurally dissimilar tasks of information re-characterization within a theory of mind context in young children, who were either assigned to card-sorting training or false-belief (perspective taking) training. The card-sorting task involved categorizing cards with two distinct features (e.g., two yellow apples, one green apple), with the relevant dimension changing (from number to color) after each set of cards was fully sorted. The false-belief task required children to answer questions about a conflicting situation in which one puppet performed an action on another, but claimed that it, instead, acted on a different puppet. To assess the training-mediated effects of card-sorting and theory of mind, two novel assessments were implemented: the card-sorting transfer task included incorporating multiple rules for new cards (sort by number then color) and sorting an entirely different set of cards on novel dimensions. The false-belief-transfer measure was a traditional Sally-Ann task using the same puppets from training. Reciprocal far-transfer was observed for both types of training – individuals receiving false-belief training improved on card-sorting, and those trained on card-sorting showed benefits on the Sally-Ann task – suggesting the presence of a shared object re-description process. Note that a similar card-sorting task resulted in transfer to "task-switching" measures in a report of near-transfer highlighted earlier (Karbach and Kray, 2009). Both sets of results point to the malleability of EFs important for perspective taking, namely, object re-description (given by the Kloo and Perner findings) and task-switching (consistent with Karbach and Kray's work). To this end, task-switching ability is apt to overlap with conflict resolution (object re-description), as switching between multiple rules involves overriding old features and rules in favor of newly relevant ones, a type of information re-characterization that is a hallmark of conflict resolution. A carefully designed training regimen – for example, by comparing task-switching training with conflict-resolution training – may illuminate the overlapping contributions of each EF for each false-belief and perspective-taking tasks similar to those outlined above.

VERBAL FLUENCY

Theory

During language production, the ease with which a lexical item is generated depends partly on the degree of competition from other candidate words. Competition demands are particularly high when multiple semantically related words are equally plausible contenders for selection (a classic case of underdetermined representational conflict; see above discussion). Items with high versus low name-agreement, for instance, present different levels of conflict during naming tasks, such that low name-agreement items associated with many alternative labels (e.g., couch/sofa/loveseat) elicit more competition, reflected by longer naming latencies, thus requiring the use of VLPFC-mediated conflict resolution to select among the competing alternatives (Kan and Thompson-Schill, 2004; Novick et al., 2009). High name-agreement items (e.g., images that invoke a single label, like apple), by contrast, have fewer alternative labels to choose from, rendering them easier to access and produce, and thus, less dependent on conflict-resolution processes. Furthermore, selection costs are compounded when cases of high-competition (low name-agreement) are crossed with increased retrieval demands (e.g., low association-strength between a cue and its most accessible response), such that items with multiple weak associates are most difficult to output (Snyder et al., 2010).

This high- vs. low-name-agreement asymmetry has been examined in non-fluent aphasic patients with VLPFC damage – the same patients mentioned above who exhibit generally poor conflict resolution and cognitive control on a variety of non-linguistic conflict-resolution tasks like Stroop and the recent-no task. This population demonstrates exaggerated effects of production difficulty for high-competition conditions that require the recruitment of conflict-resolution resources, such that they take significantly longer or even fail to produce these items altogether relative to low-competition items (Novick et al., 2009). Patients with this neuroanatomical profile have difficulty with other verbal fluency tasks, including completing sentences when the options are open-ended (and therefore ambiguous), vs. when the to-be-completed fragments provide a highly constrained context, yielding little

competition from possible alternative continuations (Robinson et al., 1998, 2005). Similarly, healthy speakers take longer to produce the names of pictured objects when they are presented in semantically homogeneous (e.g., snake, cow, dog, ant) vs. mixed contexts (e.g., snake, bus, axe, chair) due to the increase in lexical-semantic competition among semantically related competitors (Belke et al., 2005). In one study, non-fluent aphasics with circumscribed VLPFC damage generated more errors when naming objects in homogeneous contexts; a companion neuroimaging experiment further showed that even healthy adults with a greater VLPFC response to naming under homogeneous conditions are prone to more naming errors compared to individuals with less VLPFC activation (Schnur et al., 2009).

Hypothesis

Careful consideration of the literature suggests that language production under conditions of conflict appears to be modulated by general EF abilities, like those governing conflict resolution on Stroop-like tasks. Consequently, training tasks tapping these same underlying neural networks may, hypothetically, be drawn on as tools to boost word selection abilities under elevated conflict-resolution demands. The idea is that better conflict-resolution skills acquired through training might generalize to an increased ability to resolve among semantically related lexical items that compete for selection, carrying important implications for clinical interventions in populations with deficits in verbal fluency that accompany a more general deficit in conflict resolution.

Furthermore, training may also have consequences for selecting among competing alternative names during states of elevated anxiety. Indeed, one study reveals that more anxious individuals (evaluated by a composite score of anxious apprehension) are impaired relative to less anxious subjects when they must generate an associated verb (in response to a given noun) under high retrieval demands, an effect mediated by VLPFC (Snyder et al., 2010). This suggests that EF resources are depleted in cases of anxiety (Gray et al., 2002), which can negatively affect word selection processes under elevated EF demands (e.g., high-competition items). Future research on conflict-resolution training, therefore, might also address whether the right interventions can be used to offset such effects of anxiety and other deleterious affective states in both production and comprehension (but see Beilock and Carr, 2005).

SUMMARY, CAVEATS, AND FUTURE DIRECTIONS

Overall, we reviewed a sample of language tasks that depend heavily on posterior regions of left VLPFC, which support conflict-resolution abilities in a variety of populations. Among these measures there is great overlap in the EF processes involved to carry them out successfully, whether it means employing conflict-resolution to produce the right word, resolve lexical ambiguities, take a speaker's perspective to avoid errors in interpretation despite referential ambiguity, or recover from temporary misanalysis during sentence parsing. We believe that in view of these convergent findings, the theory that conflict resolution and cognitive-control contributes to language use may lead to the hypothesis that these domain-general conflict-control processes could be the target of extensive training regimens, the result of which could be attenuated processing difficulty during language use across a range of tasks, as indexed through measures of far-transfer. Such hypotheses are motivated also by the demonstration of positive transfer effects in non-linguistic cognitive domains following regimens targeting other EFs. This work could be particularly applicable to patients with lesions restricted to left posterior VLPFC, to determine (a) if their conflict-resolution performance changes on linguistic and non-linguistic tasks post-training, and (b) what new compensatory processes or brain systems they engage to support any observed performance increases (evaluated through pretest/posttest neuroimaging). There are similar implications for young children, whose comprehension might fail for similar reasons as the patients (i.e., deficits in cognitive control). Generally, this research program could suggest new inferences about the plasticity of the mind and brain, with respect to language processing especially, and the causal effects of language and cognition interactions.

Given prior evidence for far-transfer from WM training tasks to other measures such as task-switching, updating, and general fluid intelligence, the major goal that we are outlining, based on our theory of the role of left VLPFC and cognitive control in language processing, would be to design training studies in search of generalized effects to language measures, in hopes of mitigating difficulties under certain production and comprehension conditions during everyday language use. Except for a study conducted by our group on the effects of conflict-resolution training on syntactic ambiguity resolution, we are unaware of other research investigating whether broader improvements might be observed in language processing assessments in adults, both healthy and impaired, and even in young children. EF interventions might be particularly attractive in clinical arenas as a technique to remediate conflict-resolution deficits broadly construed, including how such impairments affect non-fluent production and comprehension difficulties under high-EF demands. Considering the patterns we reviewed suggesting a shared role for domain-general conflict-resolution processes across a variety of language processing tasks, a common training regimen targeting this EF could, hypothetically, be successful in correcting problems observed in each of these tasks. Future research should test this, perhaps through various ways to evaluate transfer, including behavioral changes, changes in brain-activation patterns in regions commonly recruited across training and transfer tasks, changes in evoked response potentials (McLaughlin et al., 2004), changes in neural connectivity (Geva et al., 2011), changes in eye-movement patterns and reading-time latencies, or any combination of these measures.

CAVEATS

There are, however, important caveats to consider. Despite several instances of successful generalization to unpracticed tasks, some reports describe research efforts failing to observe transfer. One explanation for the absence of transfer findings may be that in at least one study, EF training was implemented casually, rather than consistently enough to actually tax trainees' EF abilities throughout the regimen (Owen et al., 2010). In this report, not all individuals in the training group received the same exposure to training, a "dosage-dependent" factor known to confer varying

levels of transfer (Jaeggi et al., 2008). Another reason for failure to show transfer involves the use of performance-*non*-adaptive training tasks (regimens that maintain a constant level of difficulty, rather than keeping participants on the threshold of their best performance), despite strong evidence favoring such designs to facilitate transfer effects (Klingberg et al., 2005; Brehmer et al., 2011). Clearly more research is needed to determine what characterizes an appropriate training regimen, as well as how dependent transfer effects are on the amount of training an individual receives (Jaeggi et al., 2008, 2011). Finally, studies failing to show transfer might lack appropriate linking hypotheses between the types of EF required to perform certain tasks; these must be understood in order to design effective training regimens, which will ultimately inform how future intervention studies are implemented.

Furthermore, there appear to be important individual differences in training success (Chein and Morrison, 2010; Jaeggi et al., 2011), such that only certain individuals achieve performance increases on the training tasks over time, and thus demonstrate transfer to unpracticed measures shown through improved performance at retest (indeed, we observed this in our own training work). It is unclear if responders and non-responders can be categorized simply by baseline EF abilities, and these differences are unlikely due to motivational factors alone (Jaeggi et al., 2011; Novick et al., submitted for publication). So, future research should address who is most likely to benefit from training, how to identify these individuals, and how training protocols should be modified or tailored to maximize transfer across a range of groups and populations (see Shipstead et al., 2012).

Another remaining question concerns the lasting effects of training. Presumably, like physical fitness conditioning, the benefits of cognitive training do not persist indelibly without continued practice, though some have demonstrated maintained benefits three to six months after training ceased (Holmes et al., 2009; Jaeggi et al., 2011; Klingberg et al., 2005). Future work should address the long-term effects of cognitive conditioning, including the advantages of giving a periodic "booster training session" to reinstate the benefits after a regimen completion.

We included young children in our brief review of the role of conflict resolution in language use to illustrate a population whose poor EF abilities yield certain language-performance failures. However, research on training this population might proceed cautiously, particularly concerning language outcomes. The reason is that the protracted development of frontal cortex – although associated with suboptimal performance on cognitive-control (and relevant language) tasks – might actually confer certain *advantages* throughout development that overshadow the drawbacks. For instance, delayed PFC development – and by extension, delayed EF abilities – may bestow a benefit to certain aspects of cognitive development such as language acquisition (as opposed to language *performance*) and creativity (Thompson-Schill et al., 2009). There may be a complex tradeoff between bottom-up (data-driven) and top-down (rule-based) thinking in young children that may promote learning and social development. Therefore, if EF training is aimed at *enhancing* cognitive-control abilities, such interventions might have negative consequences, at least temporarily, for this population. Future work should address this concern, in addition to the long-term effects of training.

Finally, research examining healthy adults and patients with neurological disorders demonstrates that EF hinges on the involvement of a widespread network that comprises both cortical (e.g., PFC, cingulate, and parietal) and subcortical (e.g., striatal) regions, clearly not just on prefrontal cortex alone (Corbetta and Shulman, 2002; Cools et al., 2007; Burgess et al., 2011; inter alia). This pattern is bolstered by training studies documenting the underlying neural signatures accompanying post-intervention differences, including increased activation of frontoparietal regions (Olesen et al., 2004); greater structural integrity evaluated by increased fiber tracts (white matter) connecting areas adjacent to intraparietal sulcus (Takeuchi et al., 2010); and an increase in the density of cortical dopamine receptors, perhaps linked to changes in striatal structures (McNab et al., 2009). Although behavioral and neuroimaging findings suggest domain-general processes in PFC that underlie cognitive-control functions across various conditions (Thompson-Schill et al., 2005), an intricate balance exists between PFC and subcortical regions that adjusts performance over different EFs (Cools et al., 2007). Such a cortical/subcortical tradeoff should be considered when choosing training and language-transfer tasks to maximize theoretical and functional-anatomical overlap, thereby increasing the prospect of transfer yield.

FUTURE DIRECTIONS

As we have highlighted, transfer might be expected only if the EFs (e.g., conflict resolution) underlying certain language tasks are targeted through training so as to affect shared processes that facilitate performance on particular language tasks (i.e., WM training tasks not involving conflict-resolution are not expected to confer transfer). Future work might continue to identify these functional-anatomical overlaps across different memory and language tasks. We believe however that there has been sufficient data accumulated to suggest a good candidate regimen targeting VLPFC-mediated conflict-resolution processes, which could affect *certain* language processing skills.

It is important to note that although we chose to focus on conflict-resolution functions given the extant data, this does not preclude the involvement of other EFs in the abovementioned language tasks (Miyake et al., 2000). To this end, the lack of mutual exclusivity of certain general cognitive processes should be considered when interpreting transfer effects within a process-specific framework, as multiple EFs might be confounded within a single training task; thus, changes in several EFs may be responsible for resultant improvements in outcome measures, a positive outcome if the goal is to show widespread transfer (e.g., Morrison and Chein, 2011; Shipstead et al., 2012). Likewise, in the examples cited above, the magnitude of the hypothesized transfer effects will likely be sensitive to the level of conflict-resolution required for each task. The amount of transfer will hinge on the degree to which underlying EFs are shared between training and assessment tasks, and this mechanistic overlap is probably influenced by both the relative involvement of a single trained EF and the extent to which other EFs are recruited in the training and outcome tasks. For example, re-characterization of representations on the high-conflict lure trials of the *n*-back task likely requires other EFs beyond just conflict resolution (e.g., monitoring, updating).

Similarly, syntactic ambiguity resolution will, of course, rely on updating processes in addition to conflict resolution. Methodologically confounding EFs is an issue that plagues training studies, rendering it difficult to extricate distinct mechanisms entirely; however, by having linking hypotheses, EF overlap across tasks maximizes chances of successful transfer. Careful design of training regimens, including tasks performed by comparison groups – for instance by maintaining minimal task differences between training and active-control tasks (e.g., a group completing the *n*-back task without the lure component) – can help elucidate the contribution of distinct EFs.

Correspondingly, the transfer conditions under which selective improvement is observed within an assessment task may mark those relying most on the trained EF. To maximize transfer, it is important to pinpoint the measures in the assessments that capture cognitive processes of interest. For instance, in our training experiment, we argued that the strongest indices of reinterpretation ability and real-time reanalysis respectively were accuracy to comprehension questions gaging lingering effects of misinterpretation and regression-path reading time in disambiguating sentence regions. Likewise, decreased gaze duration to privileged items in a common-ground assessment task, for example, probably involves information re-characterization, rendering this a candidate measure to observe conflict-resolution training-related changes. The ability to make specific predictions for when and where transfer is selectively expected, as well as the conditions under which it is not, will ultimately lend important insight to the EFs affected during successful intervention when transfer effects are observed in studies carried out under proper linking assumptions and within a theoretically guided process-specific account.

Also worth mentioning is the contribution of several – perhaps even overlapping – domain-general resources that may be recruited during language tasks not discussed here (e.g., mnemonic aspects of WM, maintenance, updating, task-switching, etc). This should be carefully considered upon designing outcome language assessments that will be the target of transfer benefits. In fact, we strongly believe that verbal WM "span" processes, which involve maintenance, processing, and temporary storage components, must play a role in spoken language comprehension tasks in which the listener cannot review the input (as she can in normal reading) once it is spoken, without using mnemonic rehearsal strategies. This is likely true regardless of the presence of ambiguity or conflicting representations, and, indeed, verbal WM by itself has been shown to play a role in reading studies using a moving-window paradigm that does not permit rereading (Fedorenko et al., 2006). Thus, in future work it will be important to design training protocols using tasks that maximize a theoretical match between the cognitive (and neural) processes involved in assessment and training measures, including WM tasks that do not necessarily involve the conflict-resolution aspect of cognitive control, when appropriate.

Cognitive training may also provide a novel approach to understanding whether EFs are critical for a multitude of language uses. The degree to which training improvement predicts changes in language processing can reveal the EFs involved in each condition; if no transfer is observed in selective cases, one might conclude that the trained EFs do not significantly contribute to the processing of the particular language condition. This type of approach provides a powerful tool for choosing among several explanations for the same data set, where the best account of the data can be gleaned from the results of a well-designed training study that poses process-specific linking hypotheses. For example, some argue that the difficulty experienced while comprehending the meaning of abstract (compared to concrete) words hinges almost entirely on domain-general processes (Hoffman et al., 2010), while other accounts posit little to no contribution from EFs (Barsalou and Wiemer-Hastings, 2005; Rodríguez-Ferreiro et al., 2011). The opportunity exists, then, to investigate whether successful EF training permits better abstract-meaning selection.

Finally, it is important to consider a growing body of research demonstrating that balanced bilinguals enjoy certain cognitive advantages relative to their monolingual peers, as this work has important implications for language education and intervention. On tasks requiring cognitive control, some findings suggest that bilinguals outperform monolinguals selectively on trials inducing conflict across a range of tasks such as the Simon task (Bialystok et al., 2004). Other data patterns reveal a broader effect, namely that bilinguals are better at conflict *monitoring:* they perform faster on both conflict and non-conflict trials under high, but not low, conflict-monitoring conditions, in which subjects cannot predict when a conflict-related item type (an incongruent flanker trial) might occur because their appearance is equally probable relative to non-conflict trials (Costa et al., 2009). Regardless of the specifics, it has become increasingly clear that rich linguistic experience (akin to the rich cognitive experience achieved through training) benefits conflict-resolution and cognitive-control performance widely, perhaps due to bilinguals' consistent switching across the two language systems they know and/or their frequent suppression of one lexicon/grammar over another, thus placing a "premium" on EFs associated with updating, conflict resolution, and set-shifting (Martin-Rhee and Bialystok, 2008; Costa et al., 2009). In other words, lifelong bilingualism may be a naturalistic form of cognitive-control training. Indeed, future work should attempt to disentangle the various processing demands that are associated with being a bilingual speaker (e.g., frequent code switches) that might yield the putative cognitive-control advantage they show; such an understanding might help extract the various EFs, in addition to conflict resolution, that are at the heart of bilinguals' benefit. It will also be beneficial to know how bilinguals' cognitive-control advantage concerning conflict resolution or conflict monitoring influences this group's linguistic abilities on the conflict-related language tasks reviewed in this paper. For instance, does bilinguals' cognitive-control advantage result in a better ability to recover the correct interpretation of garden-path sentences, following a misanalysis? The answer to this question could suggest important inferences one could draw about the prospective impact that process-specific conflict-resolution training might have on this group.

Recent findings suggest that bilingualism confers protective benefits against cognitive decline: bilingual patients diagnosed with Alzheimer's disease (AD), who are matched on a range of factors (e.g., degree of cognitive impairment, symptomatic

expression, demographic variables) to monolinguals with the same diagnosis, have significantly *more* brain atrophy in areas commonly examined to differentiate AD patients from healthy adults (Schweizer et al., 2011). The implication is that bilinguals may have greater "cognitive reserve" than would be predicted given the amount of neuropathology they exhibit; that is, the cognitive symptoms associated with AD may be delayed in this population because of their premorbid advantage. What about bilingual children and VLPFC patients? Are they "inoculated" from the cognitive-control deficits they are otherwise known for (in monolinguals) in terms of their non-linguistic and language processing abilities under high-conflict demands? If so, what behavioral mechanisms and neural systems do they recruit to compensate?

Furthermore, will cognitive-control training over the long-term yield similar protective benefits in monolinguals? Will their performance begin to approach that of (untrained) bilinguals? Will EF training confer comparable protection against normal age-related cognitive decline (Richmond et al., 2011), regardless of AD? These are open empirical questions and might be the focus of future longitudinal research. Also: to what extent does proficiency level matter in adults who have learned a second language, regarding the cognitive-control benefits they reap and the implications for intervention? Balanced bilinguals, as sketched above, enjoy certain advantages; presumably highly proficient (but unbalanced) bilinguals and those with lower proficiency levels will pattern somewhere in between the balanced group and the monolinguals regarding cognitive-control performance, depending on the relative processing demands associated with their proficiency levels. Where such bilinguals pattern can provide useful insight into the design of future training studies to bring these groups' performance ranges closer to approximate the balanced population. How much room is there for balanced bilinguals to gain from EF training? If a highly proficient group shows a similar cognitive-control advantage to that of bilinguals, then it may suggest the prospect of similar benefits (in terms of effect sizes) gained from training. Conversely, if a low-proficiency group that rarely switches between linguistic systems does not demonstrate a cognitive-control advantage compared to monolinguals, this would suggest opportunity for EF training to bestow benefits. If neither high- nor low-proficiency groups demonstrates a cognitive-control advantage, then perhaps learning a second language in adulthood does not enhance EF abilities similar to how early acquisition of two linguistic systems does. EF training could therefore be beneficial to unbalanced groups across a range of proficiency levels. Ultimately, future work in this area will clarify our understanding of the interplay between bilingualism, cognitive control, and the effects of training on language and other tasks that share cognitive processes.

CLOSING REMARKS

EF training holds promise to result in gains in cognition and language use in both production and comprehension domains, easing processing difficulty when multiple active and equally compelling representations are at odds (underdetermined representational conflict), or when dominant biases must be reined-in (prepotent conflict). Such interventions could potentially mitigate problems in language use under generally high conflict demands, not just in special populations (e.g., non-fluent aphasics with conflict-resolution deficits), but also in healthy individuals, including developing children, who experience occasional difficulty in reading, listening, or speaking due to heightened demands for cognitive control (in some cases perhaps due to resource depletion). Such research, provided reliable demonstrations of far-transfer, would add insight to our current understanding of how broad, non-linguistic cognitive abilities contribute to language use.

ACKNOWLEDGMENTS

This research was funded by the University of Maryland Center for Advanced Study of Language and an NSF IGERT Grant on Biological and Computational Foundations of Language Diversity (grant DGE-001465).

REFERENCES

Altmann, G. T., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264.

Badre, D., and Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* 45, 2883–2901.

Balota, D. A., and Faust, M. E. (2001). "Attention in dementia of the Alzheimer's type," in *Handbook of Neuropsychology*, eds F. Boller, and S. F. Cappa (New York: Elsevier Science), 51–80.

Barsalou, L. W., and Wiemer-Hastings, K. (2005). "Situating abstract concepts," in *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, eds D. Pecher, and R. Zwaan (New York: Cambridge University Press), 129–163.

Bedny, M., Hulbert, J. C., and Thompson-Schill, S. L. (2007). Understanding words in context: the role of Broca's area in word comprehension. *Brain Res.* 1146, 101–114.

Beilock, S. L., and Carr, T. H. (2005). When high-powered people fail: working memory and "Choking Under Pressure" in math. *Psychol. Sci.* 16, 101–105.

Belke, E., Meyer, A., and Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *Q. J. Exp. Psychol.* 58, 667–692.

Bialystok, E., Craik, F. I. M., Klein, R., and Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychol. Aging* 19, 290–303.

Bilenko, N. Y., Grindrod, C. M., Myers, E. B., and Blumstein, S. E. (2009). Neural correlates of semantic competition during processing of ambiguous words. *J. Cogn. Neurosci.* 21, 960–975.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652.

Brehmer, Y., Rieckmann, A., Bellander, M., Westerberg, H., Fischer, H., and Bäckman, L. (2011). Neural correlates of training-related working-memory gains in old age. *Neuroimage* 58, 1110–1120.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychon. Bull. Rev.* 16, 893–900.

Burgess, G. C., Gray, J. R., Conway, A. R., and Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *J. Exp. Psychol. Gen.* 140, 674–692.

Chein, J. M., and Morrison, A. B. (2010). Expanding the mind's workspace: training and transfer effects with a complex working memory span task. *Psychon. Bull. Rev.* 17, 193–199.

Christianson, K., Williams, C. C., Zacks, R. T., and Ferreira, F. (2006). Younger and older adults' "Good-enough" interpretations of garden-path sentences. *Discourse Process.* 42, 205–238.

Cools, R., Sheridan, M., Jacobs, E. J., and D'Esposito, M. D. (2007). Impulsive personality predicts dopamine-dependent changes in fronto-striatal activity during component processes of working memory. *J. Neurosci.* 27, 5506–5514.

Copland, D. A., Sefe, G., Ashley, J., Hudson, C., and Chenery, H. J. (2009). Impaired semantic inhibition during lexical ambiguity repetition in Parkinson's disease. *Cortex* 45, 943–949.

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215.

Costa, A., Hernández, M., Costa-Faidella, J., and Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: Now you see it, now you don't. *Cognition* 113, 135–149.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., and Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science* 320, 1510–1512.

D'Esposito, M., and Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia* 37, 1303–1315.

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence processing: evidence against domain-specific working memory resources. *J. Mem. Lang.* 54, 541–553.

Feuerstein, R. (1980). Cognitive modifiability in adolescence: cognitive structure and the effects of intervention. *J. Spec. Educ.* 15, 269–287.

Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* 14, 178–210.

Friedman, N. P., and Miyake, A. (2004). The relations among inhibition and interference cogntiive functions: a latent variable analysis. *J. Exp. Psychol. Gen.* 133, 101–135.

Geva, S., Correia, M., and Warburton, E. A. (2011). Diffusion tensor imaging in the study of language and aphasia. *Aphasiology* 25, 543–558.

Gray, J. R., Braver, T. S., and Raichle, M. E. (2002). Integration of emotion and cognition in the lateral prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4115–4120.

Gray, J. R., Chabris, C. F., and Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* 6, 316–322.

Hamilton, A. C., and Martin, R. C. (2005). Dissociations among tasks involving inhibition: a single-case study. *Cogn. Affect. Behav. Neurosci.* 5, 1–13.

Hoffman, P., Jefferies, E., and Lambon Ralph, M. A. (2010). Ventrolateral prefrontal cortex plays an executive regulation role in comprehension of abstract words: convergent neuropsychological and rTMS evidence. *J. Neurosci.* 30, 15450–15456.

Holmes, J., Gathercole, S. E., and Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Dev. Sci.* 12, F9–F15.

Hussey, E., Teubner-Rhodes, S., Dougherty, M., Bunting, M., and Novick, J. (2010). Improving garden-path recovery in healthy adults through cognitive control training. *Talk Presented at the 16th Annual Conference on Architectures and Mechanisms for Language Processing*, York, UK.

Huttenlocher, P. R., and Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.* 387, 167–178.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., and Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6829–6833.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., and Shah, P. (2011). Short and long term benefits of cognitive training. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10081–10086.

January, D., Trueswell, J. C., and Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's Area: Implications for the neural basis of sentence processing. *J. Cogn. Neurosci.* 21, 2434–2444.

Jonides, J. (2004). How does practice make perfect? *Nat. Neurosci.* 7, 10–11.

Jonides, J., and Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience* 139, 181–193.

Kan, I. P., and Thompson-Schill, S. L. (2004). Effect of name agreement on prefrontal activity during overt and covert picture naming. *Cogn. Affect. Behav. Neurosci.* 4, 43–57.

Kane, M. J., and Engle, R. W. (2000). Working memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 333–358.

Karbach, J., and Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Dev. Sci.* 12, 978–990.

Khanna, M. M., and Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *Q. J. Exp. Psychol.* 63, 160–193.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., Gillberg, C., Forssberg, H., and Westerberg, H. (2005). Computerized training of working memory in children with ADHD – a randomized, controlled trial. *J. Am. Acad. Child Adolesc. Psychiatry* 44, 177–186.

Klingberg, T., Forssberg, H., and Westerberg, H. (2002). Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood. *J. Cogn. Neurosci.* 14, 1–10.

Kloo, D., and Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Dev.* 74, 1823–1839.

Li, S., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., and Lindenberger, U. (2008). Working memory plasticity in old age: Practice gain, transfer, and maintenance. *Psychol. Aging* 23, 731–742.

Martin-Rhee, M. M., and Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Biling. Lang. Cogn.* 11, 81–93.

McLaughlin, J., Osterhout, L., and Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nat. Neurosci.* 7, 703–704.

McNab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky, P., Forssberg, H., and Klingberg, T. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science* 323, 800–803.

Milham, M. P., Banich, M. T., Webb, A., Barad, V., Cohen, N. J., Wszalek, T., and Kramer, A. F. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cogn. Brain Res.* 12, 467–473.

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100.

Monsell, S. (1978). Recency, immediate recognition and reaction time. *Cogn. Psychol.* 10, 465–501.

Morrison, A. B., and Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychon. Bull. Rev.* 18, 46–60.

Nelson, J. K., Reuter-Lorenz, P. A., Sylvester, C. C., Jonides, J., and Smith, E. E. (2003). Dissociable neural mechanisms underlying response-based and familiarity-based conflict in working memory. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11171–11175

Nilsen, E., and Graham, S. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cogn. Psychol.* 58, 220–249.

Norman, W., and Shallice, T. (1986). "Attention to action: willed and automatic control of behavior," in *Consciousness and Self Regulation: Advances in Research and Theory*, Vol. 4, eds R. J. Davidson, R. Schwartz, and D. Shapiro (New York: Plenum), 1–18.

Novick, J. M., Kan, I. P., Trueswell, J. C., and Thompson-Schill, S. L. (2009). A case for conflict across multiple domains: Memory and language impairments follow damage to ventrolateral prefrontal cortex. *Cogn. Neuropsychol.* 26, 527–567.

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2005). Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. *Cogn. Affect. Behav. Neurosci.* 5, 263–281.

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2010). Broca's area and language processing: evidence for the cognitive control connection. *Lang. Linguist. Compass* 4, 906–924.

Olesen, P., Westerberg, H., and Klingberg, T. (2004). Increased prefrontal and parietal brain activity after training of working memory. *Nat. Neurosci.* 7, 75–79.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., and Ballard, C. G. (2010). Putting brain training to the test. *Nature* 465, 775–778.

Persson, J., Welsh, K. M., Jonides, J., and Reuter-Lorenz, P. A. (2007). Cognitive fatigue of executive processes: interaction between interference resolution tasks. *Neuropsychologia* 45, 1571–1579.

Richmond, L., Morrison, A., Chein, J., and Olson, I. (2011). Working memory training and transfer in older adults. *Psychol. Aging* 26, 813–822.

Robinson, G., Blair, J., and Cipolotti, L. (1998). Dynamic aphasia: An

inability to select between competing verbal responses? *Brain* 121, 77–89.

Robinson, G., Shallice, T., and Cipolotti, L. (2005). A failure of high level verbal response selection in progressive dynamic aphasia. *Cogn. Neuropsychol.* 22, 661–694.

Rodríguez-Ferreiro, J., Gennari, S. P., Davies, R., and Cuetos, F. (2011). Neural correlates of abstract verb processing. *J. Cogn. Neurosci.* 23, 106–118.

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., and Thompson-Schill, S. L. (2009). Localizing interference during naming: convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Proc. Natl. Acad. Sci. U.S.A.* 106, 322–327.

Schweizer, T. A., Ware, J., Fischer, C. E., Craik, F. I., and Bialystok, E. (2011). Bilingualism as a contributor to cognitive reserve: evidence from brain atrophy in Alzheimer's disease. *Cortex* 12, 8–15.

Shipstead, Z., Redick, T. S., and Engle, R. W. (2010). Does working memory training generalize? *Psychol. Belg.* 50, 245–276.

Shipstead, Z., Redick, T. S., and Engle, R. W. (2012). Is working memory training effective? *Psychol. Bull.* PMID: 22409508. [Epub ahead of print].

Smith, E. E., and Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science* 283, 1657–1661.

Snyder, H. R., Hutchison, N., Nyhus, E., Curran, T., Banich, M. T., O'Reilly, R. C., and Munakata, Y. (2010). Neural inhibition enables selection during language processing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 16483–16488.

Staub, A., and Rayner, K. (2007). "Eye movements and on-line comprehension processes," in *The Oxford Handbook of Psycholinguistics*, ed. G. Gaskell (Oxford: Oxford University Press), 327–342.

Takeuchi, H., Sekiguchi, A., Taki, Y., Yokoyama, S., Yomogida, Y., Komuro, N., Yamanouchi, T., Suzuki, S., and Kawashima, R. (2010). Training of working memory impacts structural connectivity. *J. Neurosci.* 30, 3297–3303.

Tanenhaus, M. K. (2007). "Eye movements and spoken language processing," in *Eye Movements: A Window on Mind and Brain*, eds R. P. G. van Gompel, M. H. Fischer, W. S., Murray, and R. L. Hill (Oxford: Elsevier), 309–326.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634.

Thompson-Schill, S. L., Bedny, M., and Goldberg, R. F. (2005). The frontal lobes and the regulation of mental activity. *Curr. Opin. Neurobiol.* 15, 219–224.

Thompson-Schill, S. L., Ramscar, M., and Chrysikou, E. G. (2009). Cognition without control: when a little frontal lobe goes a long way. *Curr. Dir. Psychol. Sci.* 18, 259–263.

Thompson-Schill, S. L., Swick, D., Farah, M. J., D'Esposito, M., Kan, I. P., and Knight, R. T. (1998). Verb generation in patients with focal frontal regions: a neuropsychological test of neuroimaging findings. *Proc. Natl. Acad. Sci. U.S.A.* 95, 15855–15860.

Trueswell, J. C., Sekerina, I., Hill, N. M., and Logrip, M. L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition* 73, 89–134.

Van der Linden, D., Frese, M., and Meijman, T. F. (2003). Mental fatigue and the control of cognitive processes: effects on preservation and planning. *Acta Psychol. (Amst)* 113, 45–65.

Vuong, L. C., and Martin, R. C. (2011). LIFG-based attentional control and the resolution of lexical ambiguities in sentence context. *Brain Lang.* 116, 22–32.

Weighall, A. R. (2008). The kindergarten path effect revisited: children's use of context in processing structural ambiguities. *J. Exp. Child. Psychol.* 99, 75–95.

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Ostensson, M. L., Bartfai, A., and Klingberg, T. (2007). Computerized working memory training after stroke – a pilot study. *Brain Inj.* 21, 21–29.

Ye, Z., and Zhou, X. (2009). Conflict control during sentence comprehension: fMRI evidence. *Neuroimage* 48, 280–290.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2011; accepted: 02 May 2012; published online: 21 May 2012.

*Citation: Hussey EK and Novick JM (2012) The benefits of executive control training and the implications for language processing. Front. Psychology **3**:158. doi: 10.3389/fpsyg.2012.00158*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 01 November 2011
doi: 10.3389/fpsyg.2011.00275

A cognitive architecture for the coordination of utterances

Chiara Gambi* and Martin J. Pickering

Department of Psychology, The University of Edinburgh, Edinburgh, UK

Edited by:
Christoph Scheepers, University of Glasgow, UK

Reviewed by:
Giovanni Pezzulo, National Research Council of Italy, Italy
Simon Garrod, University of Glasgow, UK

****Correspondence:***
Chiara Gambi, Department of Psychology, The University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK.
e-mail: c.gambi@sms.ed.ac.uk

Dialog partners coordinate with each other to reach a common goal. The analogy with other joint activities has sparked interesting observations (e.g., about the norms governing turn-taking) and has informed studies of linguistic alignment in dialog. However, the parallels between language and action have not been fully explored, especially with regard to the mechanisms that support moment-by-moment coordination during language use in conversation. We review the literature on joint actions to show (i) what sorts of mechanisms allow coordination and (ii) which types of experimental paradigms can be informative of the nature of such mechanisms. Regarding (i), there is converging evidence that the actions of others can be represented in the same format as one's own actions. Furthermore, the predicted actions of others are taken into account in the planning of one's own actions. Similarly, we propose that interlocutors are able to coordinate their acts of production because they can represent their partner's utterances. They can then use these representations to build predictions, which they take into account when planning self-generated utterances. Regarding (ii), we propose a new methodology to study interactive language. Psycholinguistic tasks that have traditionally been used to study individual language production are distributed across two participants, who either produce two utterances simultaneously or complete each other's utterances.

Keywords: coordination, joint action, prediction, shared representations

INTRODUCTION

The interactive use of language in conversation is a form of joint activity, in which individuals act together to achieve the common goal of communicative success. Clark (1996, 2002) proposed that conversation shares fundamental features with other joint activities, for example waltzing, playing a duet, or shaking hands. The most central, defining feature of all joint activities is coordination: the mutual process by which actors take into account the intentions and the (performed or to-be-performed) actions of their partners in the planning and performance of their own actions (Clark, 1996, pp. 61–62). Clark regards the process by which individual actors manage to coordinate to be a form of problem solving, and his focus is on "strategies" that they use to attain coordination. Despite the recognition that co-actors need to coordinate both on content (the common intended goal) and on processes ("the physical and mental systems they recruit in carrying out those intentions"; Clark, 1996, p. 59) to succeed in a joint action, very little is in fact said about such processes. To illustrate this point, we look at two aspects of coordination in language use: the synchronization of the processes of production and comprehension, and turn-taking.

First, production and comprehension never occur in isolation, but the speaker's act of production unfolds while the listener comprehends it. In order to reach mutual understanding, they need to process linguistic (and non-linguistic) signals as they occur, while monitoring for errors and misunderstandings, and usually compensating for a fair amount of noise present in the environment. Clark (1996, 2002) argued that speaker and listener synchronize their acts of production and comprehension by striving to comply with principles such as the continuity principle, which states that constituents should be produced fluently whenever possible (Clark and Wasow, 1998). When they have to deviate from these principles, they follow conventional strategies to help their listeners by signaling that one of the principles is being violated. For example, Clark (2002) assumes that speakers produce certain types of disfluencies to inform listeners that they are violating the continuity principle. But he is silent on the mechanisms that normally allow synchronization, merely pointing out that the listener needs to attend to a speaker's productions.

Second, speakers and listeners take turns by repeatedly switching roles in the conversation. This alternation is managed "on the fly" by the participants themselves, at least in informal conversations (Sacks et al., 1974; Clark, 1996). Transitions are so smooth that the average gap between turns ranges from approximately 0 ms to around 500 ms (De Ruiter et al., 2006; Stivers et al., 2009), depending on language and culture. This tight temporal coordination is coupled with coordination at the pragmatic level, since each contribution normally constitutes an appropriate response to a previous contribution by the other speaker. Coordination is thought to result from the application of a set of norms, which govern turn transitions and state who can claim the ground and when (Sacks et al., 1974). It is also recognized that the listener anticipates the end of the speaker's turn (Sacks et al., 1974; Clark, 1996, 2002); additionally, the listener starts planning her utterance in advance, while the previous speaker's turn is still unfolding.

A widespread claim in the literature on turn-taking is that speakers help their addressees by signaling whether they want to keep the floor or are about to end their turn (Clark, 2002). Many

linguistic (e.g., pitch contour) and non-linguistic (e.g., breathing) cues are reliably associated with turn-holding or turn-yielding points in a conversation. However, very few studies have systematically investigated which features of the speech signal are actually exploited by listeners to discriminate between end-of-turn and turn-holding points (see Gravano and Hirschberg, 2011; Hjalmarsson, 2011) and even fewer studies have looked at listeners' ability to use such cues on-line to anticipate turn endings (Grosjean and Hirt, 1996; De Ruiter et al., 2006; Magyari and De Ruiter, 2008). Moreover, no mechanisms have been proposed to explain how listeners can simultaneously comprehend what the speaker is saying, use the available cues to predict when the speaker's turn is going to end, and prepare their own contribution.

Another important approach to conversation as a joint activity has developed the study of coordination from a quite different perspective. Two conversational partners tend to unconsciously coordinate their body postures (Shockley et al., 2003) and gaze patterns (e.g., Richardson and Dale, 2005; see Shockley et al., 2009). One way of explaining such findings is based on the properties of oscillators, systems characterized by a periodic cycle. Mechanical oscillators (e.g., pendulums) tend to spontaneously attune their cycles, so that they become entrained: their cycles come into phase (or anti-phase). Neural populations firing at certain frequencies might act as oscillators, and sensory information regarding the phase of another oscillator (e.g., in another human body) could serve to fine-tune them. The entrainment of oscillators is therefore an automatic coordinative mechanism. According to this account, coordination, in the form of synchronization, emerges from the interaction of two dynamic systems, without any need for intentions. This view therefore suggests that coordination need not be goal-directed (Richardson et al., 2005; Shockley et al., 2009; Riley et al., 2011).

The entrainment of oscillators might explain the remarkable timing skills shown by language users. Wilson and Wilson (2005) proposed that such entrainment accounts for speakers' ability to avoid gaps or overlaps in conversation. In their account, the production system of a speaker oscillates with a syllabic phase: the readiness to initiate a new syllable is at a minimum in the middle of a syllable and peaks half a cycle after syllable offset. They argued the interlocutors converge on the same syllable rate, but their production systems are in anti-phase, so that the speaker's readiness to speak is at minimum when the listener's is at a maximum, and *vice versa*. Cummins (2003, 2009) found that two people can read the same text aloud with almost perfect synchrony; his participants only reviewed the text once and, even without any practice, could easily maintain average lags as short as 40–60 ms (Cummins, 2003). This timing is impressive, considering the huge amount of variability in speech, even within one speaker. Cummins (2009) tentatively suggested that the production systems of synchronous readers become entrained.

However, the oscillator model cannot fully explain turn-taking. First, regularities in speech appear to take place over very short time-scales, with the cyclic pattern of syllables that Wilson and Wilson (2005) propose as the basis for entrainment occurring at 100–150 ms. If predictions were made on the basis of syllable-level information alone, there would simply be not enough time to prepare the next contribution and leave a 0-ms gap. Anticipation of the end of a turn, instead, must draw on information that spans units larger than the syllable. Thus there must be additional mechanisms underlying coordination between interlocutors. In addition, Wilson and Wilson's account cannot explain how entrainment of oscillators might lead to mutual understanding.

More generally, accounts within this framework can only explain instances of rhythmic, highly repetitive activities. As such, they have no explanation for the pragmatic link between two complementary actions, be they turns in a conversation or the acts of handing over a mug and pouring coffee in it. Consider, for example, how answers complement questions. For an addressee to produce an appropriate answer, it is not enough to talk in anti-phase with the speaker. She must be able to plan in advance not only *when* to start speaking, but also *what* to say (Sebanz and Knoblich, 2009; Vesper et al., 2010).

Clark's (1996, 2002) approach and the entrainment of oscillators clearly deal with separate levels of analysis. Clark describes the dynamics of coordination at what we might call the "intentional" level. Interlocutors coordinate by making inferences about the intentions underlying their partners' behavior. Ultimately, coordination is successful if they develop mutual beliefs about their intentions. In this, they are helped by the existence of conventions (e.g., turn-allocation norms) that map intentions onto behavior. On the other hand, the entrainment-of-oscillators approach focuses on the behavioral patterns exhibited by two coordinating systems. It maintains that very general physical principles can explain the emergence of such patterns. Importantly, recent reviews (Knoblich et al., 2011) and computational accounts (Pezzulo and Dindo, 2011) have emphasized that successful joint action is likely to require coordination at both a higher level (intentions) and a lower level (bodily movements). We argue that one needs an intermediate level of analysis. In essence, it is at this level that one can define a cognitive architecture for coordination. This should comprise a set of mechanisms (representations and processes acting on those representations) that underlie coordination and, ultimately, mutual understanding between interlocutors.

In this paper, we propose that the most promising way of identifying these mechanisms stems from a mechanistic account of language processing. This is of course what psycholinguistic theories have traditionally tried to develop. However, most of these theories are concerned with monolog, in which speakers and listeners act in isolation. Pickering and Garrod (2004) pointed out the need for a theory of dialog that can explain the seemingly effortless, automatic nature of conversation. They proposed that interlocutors come to a mutual understanding via a process of alignment, whereby their representational states tend to converge during the course of a conversation. Alignment occurs at many different levels, including words and semantics (Garrod and Anderson, 1987), syntax (Branigan et al., 2000), and ultimately the situation model. Importantly, they argued that the simple mechanism of priming (i.e., facilitation in processing of an item due to having just processed the same or a related item) underlies such alignment. Alignment facilitates coordination (i.e., similar representational states facilitate successful interaction). In their model, therefore, coordination among interlocutors results from a mechanism of priming that is known to operate within the individual speaker's

production system and the individual listener's comprehension system.

To account for alignment between speaker and listener, Pickering and Garrod (2004) assumed representational parity between production and comprehension. Menenti et al. (2011) recently provided evidence for this assumption in an fMRI study, showing that brain areas that support semantic, lexical, and syntactic processing are largely shared between language production and language comprehension. In another fMRI study, Stephens et al. (2010) compared activation in a speaker with activation in listeners attending to the speech produced by that speaker. The speaker's and the listeners' neural activity were not only spatially overlapping, but also temporally coupled. As might be expected, areas of the listeners' brains were typically activated with some delay relative to the corresponding areas of the speaker's brain. However, some areas showed the opposite pattern: they were activated in the listener's brain before they were in the speaker's. These areas might be responsible for anticipatory processing of the sort that seems to be necessary for coordination. The size of areas showing anticipatory activity was positively correlated with listeners' comprehension performance. Interestingly, Noordzij et al. (2009) also found extensive overlap when comparing the planning and recognition of non-conventional communicative actions (e.g., moving a token to communicate its goal position on a game board). If the production and comprehension systems make use of the same representations, those representations that have just been built in comprehension can be used again in production and *vice versa*. Because interlocutors alternate between production and comprehension, their production and comprehension systems become increasingly attuned.

However, it is not certain that representational parity can by itself account for coordination in dialog. In addition to a common format for the representation of self-generated and other-generated actions (Sebanz et al., 2006a), addressees need to predict speakers' utterances (Pickering and Garrod, 2007) and make use of these predictions when producing their own utterances (Garrod and Pickering, 2009). To show this, the next section first reviews evidence that representational parity holds between perception and action. We show how perception–action links can serve as a basis for prediction of others' actions and explain how these predictions can in turn affect the planning of one's own actions. Then we apply these ideas specifically to the coordination of utterances.

As well as outlining a theoretical framework, we describe some experimental paradigms that can help answer the questions raised by this new approach. In fact, we believe that the inadequacy of the current accounts is partly due to the limitations associated with current experimental studies of dialog. These studies have traditionally looked at how coordination is achieved off-line, over quite long stretches of conversation, using measures such as changes in turn length or choice of referring expressions. Under these circumstances, time constraints are loose enough to allow for relatively slow and intentional cognitive processes to be the basis of coordination (e.g., Clark and Wilkes-Gibbs, 1986; Wilkes-Gibbs and Clark, 1992). Studies that focus on alignment have reduced the time-scale to consecutive utterances. Garrod and Anderson (1987), for example, analyzed the spatial descriptions produced during a co-operative maze game. They showed that interlocutors align locally on the method of description that they use to refer to locations in the maze. Studies of priming in dialog have systematically investigated this utterance-to-utterance alignment. Thus, Branigan et al. (2000) had participants alternate in the description of pictures and found that the addressee tends to re-use the syntactic structure of the description produced by the current speaker, in the following turn. However, this is still a relatively long time-scale.

In contrast, no study has looked at that moment-by-moment coordination that might explain how listeners and speakers synchronize and take turns with virtually no gap or overlap. We argue that the obvious way to do this would be to conduct experiments with more than one participant in which the relative timing of their contributions is carefully controlled and the relationship between their utterances is systematically varied. We would then be able to test whether aspects of others' utterances are indeed predicted and to what extent such predictions are taken into account when planning one's own utterances. Importantly, these experiments should focus on the study of mechanistic processes (rather than intentional behavior), and should in this respect be similar to the psycholinguistics of monolog.

REPRESENTING ANOTHER'S ACTIONS

The behavioral and neuroscientific literature on joint actions has investigated how actions performed by a co-actor are taken into account in the planning and performance of one's own actions (Sebanz et al., 2006a; Sebanz and Knoblich, 2009). Sebanz and colleagues have argued that acting together requires shared representations. This means that people should represent other people's actions alongside their own. In a series of experiments, they demonstrated that such representations are indeed formed and activated automatically, even when they are not relevant for one's own actions because the two participants are merely acting next to each other on alternating trials (as opposed to acting together to reach a common goal; Sebanz et al., 2003, 2005; see also Atmaca et al., 2008; Vlainic et al., 2010).

For example, when one participant is instructed to respond to red stimuli with right button presses and the other responds to green stimuli with left button presses (joint condition), reaction times are slower when the stimulus and the response are spatially incongruent (e.g., the red stimulus points to the left) than when they are congruent. A similar interference effect arises when a single participant is in charge of both responses (individual condition; Sebanz et al., 2003, 2005). In the individual condition, the irrelevant spatial feature of the stimulus automatically activates the spatially congruent response, which is part of the participant's response set. In the joint condition, there is only one response in each participant's response set. However, the partner's task is represented as well; the presentation of a leftward-pointing stimulus automatically evokes the partner's response (left button press) as well as one's own (right button press), yielding interference. Additionally, electrophysiological evidence suggests that the action associated with the partner's task is inhibited on no-go trials (Sebanz et al., 2006b). In these experiments, knowledge about the partner's task is available from the start (i.e., both participants listen while task instructions for each co-actor are given) and can be used to predict the partner's action response even when there is

no sensory feedback from the other's actions (Atmaca et al., 2008; Vlainic et al., 2010); seeing the associated stimulus is enough to activate the appropriate response (Sebanz et al., 2006a).

When knowledge about others' actions is not available as part of a task specification, the mere observation of actions performed by others can still lead to the formation of shared representations (Sebanz et al., 2006a). More precisely, the action system might be involved in action observation. At least two lines of evidence support this claim. First, observing an action that is incompatible with a planned action affects execution of that action (e.g., Brass et al., 2000; see Wilson and Knoblich, 2005); second, areas of the motor system involved in action planning are activated during passive observation of the same actions (e.g., Iacoboni et al., 1999; see Rizzolatti and Craighero, 2004 for a review). This suggests that observed actions are coded in the same format as one's own actions (Prinz, 1997; Sebanz et al., 2006a).

Many researchers agree that motor involvement in action perception can aid action understanding (e.g., Blakemore and Decety, 2001; Buccino et al., 2004). Wilson and Knoblich (2005) proposed that action perception involves *covert imitation* of others' actions, as the perceiver internally simulates the observed action in her own motor system. The simulation is quicker than the actual performance of an action. Therefore, it can also be used to formulate perceptual predictions about what the observed actor is going to do next. Such predictions allow rapid and effective interpretation of the observed movement, even in cases where the movement needs to be partially reconstructed, because perceptual information is missing (predictions would serve to "fill in the gaps"). In addition, covert imitation of the partner in a joint activity could underlie quick and appropriate reactions to his or her actions (Wilson and Knoblich, 2005, p. 468).

More specifically, Wilson and Knoblich (2005) proposed that covert imitation of others is based on a model of one's own body (cf. Grush, 2004). Though this model can be adjusted to accommodate differences between the observer's and the actor's bodies, it follows that simulation (and hence prediction) of one's own actions should be more accurate than simulation of actions performed by others. In support of this claim, people are better at predicting a movement trajectory (e.g., in dart-throwing or handwriting) when watching a video of themselves vs. others (Knoblich and Flach, 2001; Knoblich et al., 2002) and pianists find it easier to synchronize with a recording of themselves than with a recording of somebody else (Keller et al., 2007).

The model that computes predictions is specifically a forward model (Wilson and Knoblich, 2005). It takes a copy of the motor command sent to the body as input and produces the expected sensory feedback as output. Expected sensory consequences of executing a motor command (e.g., expected limb position) can then be compared with actual feedback coming from the sensory system. This mechanism allows for fast, on-line control of movements (Wolpert and Flanagan, 2001). If the actual position of a limb, for example, does not match the predicted position, adjustments can be made to the motor command to minimize the difference. When the forward model is run, activation of the motor system normally ensues. However, when the forward model is used to covertly imitate another actor, covert imitation does not always result in overt imitation of another's movements. It is likely that the overt motor response is suppressed in such cases (Grush, 2004; Sebanz et al., 2006b).

Finally, and again following Sebanz et al. (2006a), we note that representing the actions performed by others and predicting what they are going to do are necessary but not sufficient for on-line coordination. What is also required is a mechanism for integrating self-generated and other-generated actions in real time. If individual actions are coordinated to the partner's actions on a moment-by-moment basis, then other-generated actions must be considered during planning of one's own actions. In support of this, Knoblich and Jordan (2003) had participants coordinate button presses that caused a circular stimulus to accelerate either to the right or to the left (with each participant being in charge of one direction) so that the stimulus remained aligned with a moving dot. Provided that feedback about the other's actions was available, participants mastered the task as successfully as participants acting alone. In particular, they learned to jointly anticipate sudden changes in the dot's movement direction.

The authors concluded that the participants were predicting the consequences of integrating their own and their partner's actions and suggested two mechanisms that could underlie this ability. Participants might run multiple simulations corresponding to the combination of the various action alternatives available to themselves and their partners (cf. Wilson and Knoblich, 2005). The other alternative, which they favored (Knoblich and Jordan, 2003; Sebanz and Knoblich, 2009), is based on the distal coding theory (Prinz, 1997; Hommel et al., 2001), which states that actions are coded in terms of the events resulting from them. Integration of self- and other-generated actions could occur at the level of these distal events. Rather than building and constantly updating a simulation of other-generated actions, then, people would simply take into account the perceptual consequences of others' actions (the events potentially resulting from them), in the same way as they would take into account other aspects of the environment (e.g., the presence of obstacles; cf. Sebanz and Knoblich, 2009, p. 361). One would then adjust one's own action plan accordingly, so that the intended event (corresponding to the joint action goal) is realized.

To summarize, the shared representational approach maintains that (i) other-generated and self-generated actions are represented in the same format, (ii) representations of other-generated actions can be used to drive predictions, and (iii) self-generated and other-generated actions are integrated in real time to achieve coordination (Sebanz et al., 2006a). By referring to representations and processes that make use of those representations, the account provides explanations at a level that bridges purely intentional and purely mechanistic accounts of coordination. Despite the above-mentioned limitations (see Introduction), entrainment of oscillators could still play an important role in coordination. In particular, it could serve as a basis to optimize other mechanisms (Vesper et al., 2010). Recall that covert imitation of other-generated actions is assumed to exploit a model of one's own body. If some basic properties of this system, such as the frequency of rhythmic unintentional movements, become attuned *via* entrainment, then simulations of another's actions would likely become more accurate, because the simulated system will end up sharing features of the system on which simulations are based. In accord with this view, co-actors that rocked chairs in synchrony were faster at

jointly moving a ball through a labyrinth (Valdesolo et al., 2010). Therefore, entrainment with another actor can enhance performance on a subsequent, unrelated joint task. Entrained actors did feel more similar to each other and more connected, but these feelings did not predict performance. Instead, enhancement appeared to be mediated by increased perceptual sensitivity to each other's actions (Valdesolo et al., 2010).

REPRESENTING ANOTHER'S UTTERANCES

In this section, we propose that interlocutors also coordinate via three mechanisms: (i) they represent others' utterances in a similar format as their own utterances; (ii) they use these representations as a basis for prediction; and (iii) they integrate self- and other-representations on-line. Interestingly, there is plenty of evidence for a direct link between speech perception and speech production (Scott et al., 2009). Fowler et al. (2003) showed that people are faster at producing a syllable in response to hearing the same syllable than in response to a tone; in fact, shadowing a syllable yielded response latencies that were nearly as fast as those found when the to-be-produced syllable was fixed and known in advance. Moreover, Kerzel and Bekkering (2000) demonstrated an action perception compatibility effect for speech (due to a task-irrelevant stimulus). They found that participants pronounced a printed syllable while watching a video of a mouth producing the same syllable more quickly than when the mouth produced a different syllable. While the first study involves intentional imitation, the second one provides more compelling evidence for automaticity. However, they both deal with cases of *overt* imitation, where there is an overt motor response. Evidence that bears more on the issue of *covert* imitation comes from neuropsychological studies of speech perception. These studies found activation of motor areas during passive listening to speech (e.g., Wilson et al., 2004), showed that this activation is articulator-specific (Pulvermüller et al., 2006), and found that stimulation of motor areas with TMS can influence speech perception (Meister et al., 2007; D'Ausilio et al., 2009; see Pulvermüller and Fadiga, 2010).

In addition, some researchers have proposed that activation of motor areas during speech perception might reflect the dynamics of forward models. In Guenther and colleagues' model of speech production, a forward model is used to compute the auditory representation corresponding to the current shape of the vocal tract, which in turn is derived from combined proprioceptive feedback and a copy of the motor command sent to the articulators (e.g., Guenther et al., 2006). In an MEG study, Tian and Poeppel (2010) demonstrated that auditory cortex is activated very quickly (around 170 ms) when participants are asked to imagine themselves articulating a syllable. They therefore proposed that forward models involved in speech production can be decoupled from the movement of the articulators. Their findings open up the possibility that a forward model of the articulation system could be used in covert imitation of perceived speech.

Activation of motor areas during speech perception could serve a variety of purposes. First, it could help understanding, just as it may for other actions (see Representing Another's Actions). In support of this, overt imitation of an unfamiliar accent (which must of course involve activation of such areas) improves accent comprehension more than mere listening (Adank et al., 2010). Alternatively, it could reflect articulatory rehearsal in the verbal working memory system (Wilson, 2001). Scott et al. (2009) suggested that motoric activation during speech perception might also facilitate coordination between language users in dialog. In particular, they proposed that the activation of the motor system underlies synchronization of the rhythmic properties of speech (entrainment). Our proposal differs in that we claim that it could also be responsible for the covert imitation, and prediction, of others' utterances (Pickering and Garrod, 2007).

WHAT KIND OF INFORMATION IS REPRESENTED?

Consider two speakers, *A* (female) and *B* (male), producing two utterances roughly at the same time, in response to a shared stimulus, such as a to-be-named picture of a kite. **Figure 1** illustrates the range of information that *A* could represent about her own utterance (upper box) and about *B*'s utterance (lower box). Before we discuss the nature of these representations, we will briefly illustrate the time course of word production, taking *A*'s production of "kite" as an example (see the timeline at the top of **Figure 1**). Models of single word production (e.g., Levelt et al., 1999) involve at least (i) a semantic representation (sem_A) corresponding to the target concept (KITE); (ii) a syntactic representation (syn_A) – sometimes called a lemma – that incorporates syntactic information about the lexical item, such as that it is a noun ($kite_{(N)}$); (iii) a phonological representation ($phon_A$) that specifies a sequence of phonemes and its syllable structure (/kaIt/). Finally, the appropriate articulatory gestures are retrieved and executed (art_A).

Note that each processing level is characterized not only by the content of the associated representation, but also by its timing [$t(sem_A)$, $t(syn_A)$, etc.]. Some representations are typically ready before others and the processing stages take different amounts of time. Indefrey and Levelt (2004) derived indicative time windows from a meta-analysis of several word production experiments. Their estimates are also reported at the top of **Figure 1**, though the exact times might depend on the words used or the experimental conditions (cf. Sahin et al., 2009, for estimates based on intracranial electrophysiological recordings).

Now, consider the upper box of **Figure 1**. We assume that *A* can generate predictive estimates of the duration of each processing stage (indicated by $\hat{t}$ in **Figure 1**). For example, she might generate the estimate $\hat{t}\ (syn_A) \approx 250$ ms, meaning that she predicts retrieving the syntactic representation will take approximately 250 ms (from picture onset). These estimates can in turn be exploited by *A* to guide planning of her own utterance. Interestingly, some studies have shown that individual speakers can coordinate the production of two successive utterances so as to minimize disfluencies (Griffin, 2003; cf. Meyer et al., 2007). Similarly, Meyer et al. (2003) demonstrated that the amount of planning speakers perform before articulation onset can depend on the response time deadline they implicitly set for their performance at a naming task. This suggests that timing estimates are computed for one's own utterances and can be used to guide planning.

Clearly, for a speaker to be able to use the information provided by timing estimates effectively, the estimates must be ready before processing at the corresponding stages is completed. So, for instance, the estimate $\hat{t}\ (syn_A) \approx 250$ ms is useful only if it is available before syntactic processing is complete. This means that

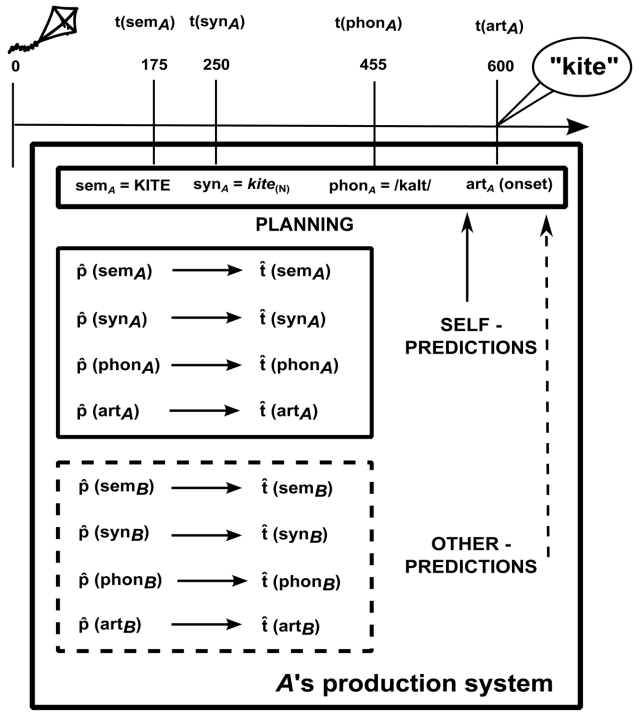


FIGURE 1 | Simultaneous production. A produces the word *kite* in response to the picture of a kite. sem$_A$, syn$_A$, phon$_A$ are semantic, syntactic, and phonological representations for *A*'s utterance. t(sem$_A$), t(syn$_A$), t(phon$_A$) indicate the actual time elapsed from picture onset (in ms) when processing is completed at each stage and the corresponding representation has been built (based on Indefrey and Levelt, 2004); t(art$_A$) marks the onset of *A*'s utterance. t̂ (sem$_A$), t̂ (syn$_A$), t̂ (phon$_A$), and t̂ (art$_A$) are timing estimates computed by *A* for her own utterance. p̂ (sem$_A$), p̂ (syn$_A$), p̂ (phon$_A$), p̂ (art$_A$) are the content predictions for *A*'s own utterance, on which the timing estimates are based. *A* believes that *B* is speaking in response to the same picture. Dotted lines refer to representations of the other. t̂ (sem$_B$), t̂ (syn$_B$), t̂ (phon$_B$), and t̂ (art$_B$) are timing estimates computed by *A* for *B*'s utterance. p̂ (sem$_B$), p̂ (syn$_B$), p̂ (phon$_B$), p̂ (art$_B$) are *A*'s content predictions, at the various processing stages, for *B*'s utterance. Horizontal arrows [from p̂ (sem$_A$) to t̂ (sem$_A$), from p̂ (syn$_A$) to t̂ (syn$_A$), etc.] indicate that estimates of the timing at each level are based on content predictions at the same level. Timing estimates at one level could also be directly based on content estimates at other levels, but we ignore this here for simplicity. Vertical arrows from self- and other-predictions to planning represent the integration stage.

the estimates are predictions. What are such predictions based on? Importantly, in language production, timing aspects are known to be closely related to the content of the computed representations. For example, word frequency affects t(phon$_A$), with phonological retrieval being slower for less frequent word forms (e.g., Caramazza et al., 2001). We therefore assume that *A* predicts aspects of the content of sem$_A$, syn$_A$, and phon$_A$. In other words, the speaker anticipates aspects of the semantics, syntax, and phonology of the utterance she is about to produce, before the representations corresponding to each level are built in the course of the production process itself. To distinguish these predictions that relate to content from predictions that relate to timing (i.e., the timing estimates), we label them p̂ (sem$_A$), p̂ (syn$_A$), p̂ (phon$_A$), and p̂ (art$_A$).

There is much evidence that content predictions of the sort we are assuming for production are indeed formulated by readers and listeners during comprehension. For example a series of sentence comprehension studies showed that predictions are made at the syntactic (lemma) level, in relation to syntactic category (e.g., Staub and Clifton, 2006) and gender (e.g., Van Berkum et al., 2005), and at the phonological level (e.g., DeLong et al., 2005; Vissers et al., 2006). For a review of some of this evidence, see Pickering and Garrod (2007), who also argued that such predictions rely on production processes. Note, however, that timing and content predictions for self-generated utterances need not always be as detailed as these studies may suggest. The specificity of predictions might depend on task demands (e.g., whether fine-grained control over the production process is needed) and be highly variable.

Having posited that predictions of timing and content can be generated for one's own utterances, we now propose that representing others' utterances can also involve the computation of predictions, and that those predictions are in a similar format to the timing and content predictions for self-generated utterances. The lower (dashed) box in **Figure 1** shows the range of information that *A* could represent about *B*'s utterance. Importantly, *A* may well not represent all of this information under all circumstances. Later, we describe experimental paradigms that can investigate the conditions under which aspects of *B*'s utterance are represented and how. Here, our aim is to provide a comprehensive framework in which such questions can be addressed.

First of all, *A* could estimate the time course of *B*'s production. Minimally, *A* could compute t̂ (art$_B$), an estimate of *B*'s speech onset latency. In addition, *A* might compute timing estimates for the different processing stages, from semantics to phonology [t̂ (sem$_B$), t̂ (syn$_B$), t̂ (phon$_B$), and t̂ (art$_B$) in **Figure 1**], just as she does when anticipating the timing of her own productions. As timing estimates are likely to be based on information regarding the content of the computed representations, we suggest that *A* can also represent the content of *B*'s utterance. In particular, *A* builds predictive representations of the semantics, syntax, and phonology of the utterance produced by *B* [p̂ (sem$_B$), p̂ (syn$_B$), p̂ (phon$_B$), and p̂ (art$_B$) in **Figure 1**].

THE NATURE OF THE REPRESENTATION OF THE OTHER

We have just proposed that other-generated utterances can be represented in a format that is similar to that of content (p̂) and timing (t̂) predictions for self-generated utterances. How are such predictions computed? We propose that people can make content and timing predictions, for both self-generated and other-generated utterances, using forward models of their own production system. This, in essence, amounts to an extension of the covert imitation account (Wilson and Knoblich, 2005) to language. Pickering and Garrod (submitted) provide a detailed theory that incorporates these claims (see also Pickering and Garrod, 2007; Garrod and Pickering, 2009).

The model is primarily used in the planning and control of one's own acts (here, speech production acts), but it can be used to simulate the production system of another speaker. When this happens, the model is decoupled from the production system, so that covertly simulating another's utterances does not lead to the actual planning of that utterance or to its articulation. In other

words, *A* does not build sem_B, syn_B, and $phon_B$ (semantic, syntactic, and phonological representations for the utterance that *B* is going to produce) just as she does not initiate art_B (the articulation stage for *B*'s utterance).

Nevertheless, speakers can overtly imitate a speaker (e.g., in speech shadowing; see Marslen-Wilson, 1973) and they sometimes complete each other's utterances (see Pickering and Garrod, 2004). On occasion, therefore, covert simulation of *B*'s utterance, *via* the computation of a forward model, results in activation of *A*'s own production system. In this case, there will be activation of the semantic (sem_B), syntactic (syn_B), and phonological ($phon_B$) representations corresponding to *B*'s to-be-produced utterance, within *A*'s production system. Depending on the predictability of *B*'s utterance, and on the speed of the simulation, *A* might end up shadowing *B*'s speech, talking in unison with *B* or even anticipating a completion for *B*'s utterance. Note, however, that some activation of *A*'s production system does not necessarily entail that *A* overtly articulates *B*'s utterance.

Note that this account differs slightly from the dominant view in the action and perception literature (e.g., Grush, 2004). According to this view, the motor system is in fact always activated following the activation of the forward model, but this activation is inhibited and therefore does not result in an overt motor response (though residual muscle activation can be detected in the periphery; e.g., Fadiga et al., 2002). The system responsible for language prediction might function in the same way as the system responsible for motor predictions. However, it is also possible that predicting *B*'s utterances does not involve any (detectable) activation flow in *A*'s language production system. At present, determining exactly under which conditions *A*'s production system is activated, and to what extent, is still a matter for empirical investigation. In the section on "Simultaneous Productions" we indicate which experimental outcomes are to be expected under the alternative hypotheses.

Another important issue relates to the accuracy of both the timing and content representations of another's utterances. For example, how similar is $\hat{p}$ (sem_B) to *B*'s concept KITE, or how accurate an estimate of *B*'s speech onset latency is $\hat{t}$ (art_B)? We expect representations of another's utterances to be generally somewhat inaccurate. First, although context and task instructions might highly constrain the productions of both speakers in experimental settings, normally *A* would have only limited information regarding what *B* intends to say. Second, *A* has limited experience of other speakers' production systems. The forward model she uses to compute predictive estimates is fine-tuned to her own production system rather than to *B*'s production system (Wolpert et al., 2003). As a consequence, timing estimates based on a model of *A*'s production system are likely to diverge from the actual time course of *B*'s production. The degree of error will also depend on how much *B* differs from *A* in speed of information processing. Conversely, we expect accuracy to increase the more *A*'s and *B*'s systems are or become similar (Wolpert et al., 2003). In conversations, the two systems might become increasingly attuned *via* alignment (Pickering and Garrod, 2004), thanks to priming channels between the production and comprehension systems of the two interlocutors. Furthermore, interlocutors' breathing patterns and speech rates can converge *via* entrainment (see Wilson and Wilson, 2005 and references therein).

Finally, we might ask whether predictions about other-generated utterances can influence the planning of one's own utterances to the same extent as predictions about self-generated utterances. For example, say that $\hat{t}$ (art_A) is a prediction of when *A* will finish articulating her current utterance. *A* should take this prediction into account as she plans when to start her next utterance. Similarly, if *B* is the current speaker and *A* wants to take the next turn, *A* could compute $\hat{t}$ (art_B), an estimate of when *B* will stop speaking. Then the question is, will *A* pay as much attention to $\hat{t}$ (art_B) as she would to $\hat{t}$ (art_A) in the first case? This is likely to depend on the circumstances. For example, $\hat{t}$ (art_B) might be weighted as less important if its degree of accuracy is low (i.e., previous predictions have proved to be wrong). Alternatively, *A* might not take $\hat{t}$ (art_B) into account, simply because she does not share a goal with *B*; for example, she might be trying hard to be rude and interrupt *B* as much as possible.

THE TIME COURSE OF PLANNING, PREDICTION, AND THEIR INTEGRATION

What is the time course of predictions, both with respect to one another and to the time course of word production? Firstly, predictions should be ready before the corresponding production representations are retrieved in the process of planning an utterance. Secondly, since we assumed that timing estimates are computed on the basis of content predictions, $\hat{p}$ (sem_A) should be ready before $\hat{t}$ (sem_A), $\hat{p}$ (syn_A) before $\hat{t}$ (syn_A), etc. Similarly for other-predictions, $\hat{p}$ (sem_B) should be ready before $\hat{t}$ (sem_B), $\hat{p}$ (syn_B) before $\hat{t}$ (syn_B), etc. (see horizontal arrows in **Figure 1**).

However, we intend not make any specific claim about the order in which predictions at the different levels (semantics, syntax, and phonology) are computed. It might be tempting to stipulate that the prediction system closely mimics the production system in this respect. In fact, however, the prediction system is a (forward) model of the production system and such a model need not implement all aspects of the internal dynamics of the modeled system. In particular, the prediction system for language could involve the same representational levels as the language production system, but the time course with which predictions are computed could differ from the time course of language production. Predictions at the levels of semantics, syntax, and phonology might even be computed separately and (roughly) simultaneously (Pickering and Garrod, 2007). In other words there could be separate mappings from the intention to communicate to semantics, syntax, and phonology. For this reason, in **Figure 1** we simply list the different predictions. Nevertheless, it is certainly the case that predictions at different levels are related to each other. For example, a prediction that the upcoming word refers to an object (a semantic prediction) and that it is a noun (a syntactic prediction) are related (because nouns tend to refer to objects). It is likely that the prediction system for language exploits such systematic relations between levels.

Once predictions are computed, how are they integrated in the process of planning an utterance (cf. vertical arrows in **Figure 1**)? To illustrate, take the following situation. The speaker needs to initiate articulation (art_A) rapidly, perhaps because of task

instructions (in an experiment) or because of an impatient listener trying to get the floor. But she also knows that her chosen word is long (e.g., helicopter). The speaker computes $\hat{p}$ (phon$_A$), a prediction of the phonology of the word. On the basis of this, the speaker estimates, $\hat{t}$ (phon$_A$), that the complete phonological representation for that word will take a long time to construct, and that she will not be able to get it ready before the timeout. The predicted failure to meet the goal either (i) causes more resources to be invested in planning to speed things up, or, if processing speed is already at limit (ii) leads to early articulation of the first syllable of the word, even if the remaining syllables have not been prepared yet (Meyer et al., 2003). In other words, predicted outcomes (i.e., the output of the forward model) can trigger corrections to the ongoing planning process, in case such outcomes do not correspond to the intended goal.

METHODOLOGICAL RATIONALE: COMPARING SELF's AND OTHER's REPRESENTATIONS

How can we test whether the proposed account is correct? First, we should identify the conditions under which other-representations are formed. Second, we should investigate the nature of such representations. To do so, we need to compare individual production and joint production (in analogy with the joint action literature; e.g., Sebanz et al., 2003). In particular, we consider two instances of joint production: simultaneous productions (see Simultaneous Productions) and consecutive productions (see Consecutive Productions). In both sections, we first introduce the rationale behind joint production tasks and present the model's general predictions. Then, we describe a few specific methods in more detail. These make use of psycholinguistic tasks that (i) have been successfully employed in the study of isolated individual production, and (ii) can be distributed between two participants to study joint production. After a brief overview of the results typically found in individual production experiments, we list the specific predictions that our account makes with regard to the comparison of the individual and the joint task in each case.

SIMULTANEOUS PRODUCTIONS

Consider two speakers planning two different or similar utterances at the same time (see **Figure 1**). If *A* automatically represents *B*'s utterance as well as her own, then her act of production will be affected by the nature of his utterance, even if there is no need for coordination; the same holds for *B*'s representation of *A*'s utterance. We therefore expect joint simultaneous production to differ from individual production. By manipulating the relationship between the two speakers' utterances (e.g., whether they produce the same or different utterances), we can further investigate the nature of *A*'s representations of *B*'s utterances.

In particular, if predictions regarding other-generated utterances are computed *via* a model of one's production system, it should be possible to simulate another's utterances *without* the corresponding representations being activated in one's own production system. Additionally, it might be possible to maintain two models active in parallel (Wolpert et al., 2003; Wilson and Knoblich, 2005), for one's own and one's partner's utterances. However, using the same format simultaneously for simulating oneself and another may well lead to competition (Hamilton et al., 2004; Wilson and Knoblich, 2005). If so, we expect greater interference from *B*'s utterance on *A*'s production when *A* and *B* perform the same act of production than when they perform different acts.

Nevertheless, if (at least partial) activation of *A*'s own production system follows her simulation of *B* *via* the forward model, then we expect representations of *B*'s utterances to interact with representations of *A*'s own utterances in the way that representations for different self-generated utterances should interact. What would be the effect of such interaction *within A*'s production system? There might be facilitation or interference, depending on a variety of factors (e.g., whether *B* is producing the same word or a different word; in the latter case, whether the two words are related in form or meaning; cf. Schriefers et al., 1990).

Besides, since some representations are harder to process than others, variables that affect processing difficulty of self-generated utterances should also exert an effect in relation to other-generated utterances. Consider, for instance, the following situation. *A* and *B* name different pictures. The frequency of picture names is varied, so that on some trials *B* produces low-frequency words, whereas on others he produces high-frequency words. Given that it is harder to access the phonological representation of a low-frequency word than a high-frequency word (cf. Miozzo and Caramazza, 2003), we predict that representing *B*'s utterance will interfere more with *A*'s naming in the low-frequency condition than the high-frequency condition. In general, the difficulty of *B*'s task will affect the degree to which the representation of *B*'s utterances affects *A*'s production of her own utterances.

To sum up, paradigms that involve two speakers' simultaneously or near-simultaneously producing utterances serve two purposes: they test whether self- and other-generated utterances are represented in the same way, and they can elucidate the nature of other-representations, and in particular whether they involve the activation of one's own production system. Below we describe two such paradigms in more detail: joint picture–word interference and joint picture–picture naming.

Joint picture–word interference

In the classical picture–word interference paradigm (individual task), naming latencies are affected by the relationship between the pictures that the participant is required to name and words superimposed on those pictures. For example, semantically related distractor words lead to longer latencies than unrelated distractor words (Schriefers et al., 1990). The task-irrelevant stimulus (word) is thought to be automatically processed and interfere with the response to the task-relevant stimulus (picture).

In a joint version of this task, participants take turns to name the picture and to perform a secondary task, which is either congruent or incongruent with the primary task of picture naming. One possibility is for the participants to be in the same room, with the congruent task being tacit naming of the picture and the incongruent task being tacit naming of the word. Alternatively, the participants could be in separate and soundproofed rooms, in which case the secondary task could be overt picture or word naming. In any case, we would have a SAME condition (congruent secondary task), in which both participants produce the same utterance (i.e., the picture's name) and a DIFFERENT condition

(incongruent secondary task), in which they produce different utterances (i.e., the picture's name and the distractor word). If speakers represent the processes underlying their partners' acts of speaking, we expect both the SAME and DIFFERENT conditions to differ from the individual task. If speakers represent the processes underlying their partners' response *via* a forward model, we expect longer latencies in the SAME than the DIFFERENT condition. If representing the other involves activation of one's own production system, on the contrary, we expect faster latencies in the SAME than in the DIFFERENT condition. In addition, we may find enhanced effects of distractor words on the processing of the pictures (e.g., greater semantic interference) in the DIFFERENT condition.

Joint picture–picture naming

In picture–picture naming tasks, participants name a target picture which is presented in the context of another (distractor) picture. The distractor picture is either related or unrelated to the target picture. Unlike picture–word interference experiments, picture–picture naming experiments typically show no clear effect of semantically related distractors on target naming latencies (e.g., Navarrete and Costa, 2005). In a joint version of the picture–picture naming task, participants either name one picture or remain silent. For trials on which the participant is naming a picture, we vary whether the partner remains silent (NO condition) or names the same (SAME condition) or a different picture (DIFFERENT condition). Assuming that the task-irrelevant picture's name is not automatically activated when performing the individual task, the NO condition should act as a control. If the participant represents the fact that her partner is naming a picture, then this may similarly affect both the SAME and the DIFFERENT condition; if she represents that her partner is naming a specific picture, we predict the SAME and the DIFFERENT condition will differ from each other. Again, the direction of these effects will depend on whether or not the production system is implicated in the representation of the other (see The Nature of the Representation of the Other).

CONSECUTIVE PRODUCTIONS

One concern with the study of simultaneous production is that it is comparatively rare in real conversations. Of course, speakers do occasionally contribute at the same time, for example when two listeners both claim the ground (e.g., in response to a question; Wilson and Wilson, 2005) or in intended choral coproduction (e.g., mutual greetings; Schegloff, 2000). But it may be that speakers do not need a system that is specialized for representing their own utterance and a simultaneous utterance by their partner.

In contrast, consecutive production occurs all the time in conversation. First, the norm in dyadic conversations is the alternation of speaking turns. Second, conversational analysts have noted the occurrence of "collaborative turn completion" (Lerner, 1991). As illustrated in Example 1 below, *B*'s act of production completes *A*'s act appropriately and with minimum delay (0.1 means 100 ms). Instances of "collaborative turn completion" are striking, because two people effectively coordinate to jointly deliver one well-formed utterance.

1. *A*: so if one person said he could not invest (0.1)
   B: then I'd have to wait

(Lerner, 1991, p. 445)

Thus, speakers have much more need of representing their own utterance and their partner's upcoming utterance. Consecutive production paradigms should then somewhat mimic the naturalistic situation exemplified in 1. For example, *A* and *B* could be shown two pictures (e.g., of a wig and of a carrot), one on the right and one on the left of a computer screen. *A* first names the left picture (*wig*); then *B* names the right picture (*carrot*; see **Figure 2A**). They are told to minimize delay between the two names (cf. Griffin, 2003). We therefore create a joint goal for them. This situation certainly differs from naturally occurring instances of "collaborative turn completion", but it allows clear experimental control, and is arguably comparable to using tasks such as picture naming to understand natural monolog. (In an alternative version of the task, participants might simply start speaking in response to cues, which might occur at different times (i.e., SOAs) depending on condition.)

Figure 2A presents a schematic description. Given the complexity of the situation, in order to ensure that the figure is readable, we illustrate what happens from the perspective of *A*, the speaker that names the first picture. The timeline at the top shows the time course of word production for *A*'s utterance (and the onset of *B*'s utterance). Just as for the simultaneous production paradigm, we assume that *A* generates timing estimates for her own utterance and that these estimates are based on content predictions (left box). In addition, we hypothesize that *A* represents *B*'s upcoming utterance in a similar format and computes timing estimates and content predictions for that utterance, as well (right box).

To test these hypotheses, we again compare joint tasks with solo tasks. In the solo task (see **Figure 2B**), which was first used by Griffin (2003), *A* produces both pictures' names, with the same instruction of avoiding pausing between the two. Clearly, *A* goes through all the processing levels for both words and builds representations at each level. The timeline at the top of panel B differs from the one in **Figure 1**: most notably, $t(art_A)$ corresponds to 1200 ms, instead of the 600 ms posited by Indefrey and Levelt (2004). This reflects the finding that participants tend to delay the onset of the first word, presumably because they perform advance planning. They start planning the second word before initiating the articulation of the first one (Griffin, 2003). We also assume that *A* computes timing estimates and content predictions for the second word, as well as for the first word.

If content and timing predictions computed for *B*'s utterance in the JOINT condition are similar to those computed for *A*'s own second utterance in the SOLO condition, we expect the JOINT and the SOLO condition to show similar patterns of results. Of course, we might also expect any effects to be weaker in the JOINT than in the SOLO condition, if other-representations are weighted less than self-representations (see The Nature of the Representation of the Other). We know that the amount of planning that speakers perform before articulation onset (and, consequently, speech onset latency) depends on various properties of the planned material, such as its length (Meyer et al., 2003) or syntactic complexity (e.g., Ferreira, 1991). Therefore, we expect

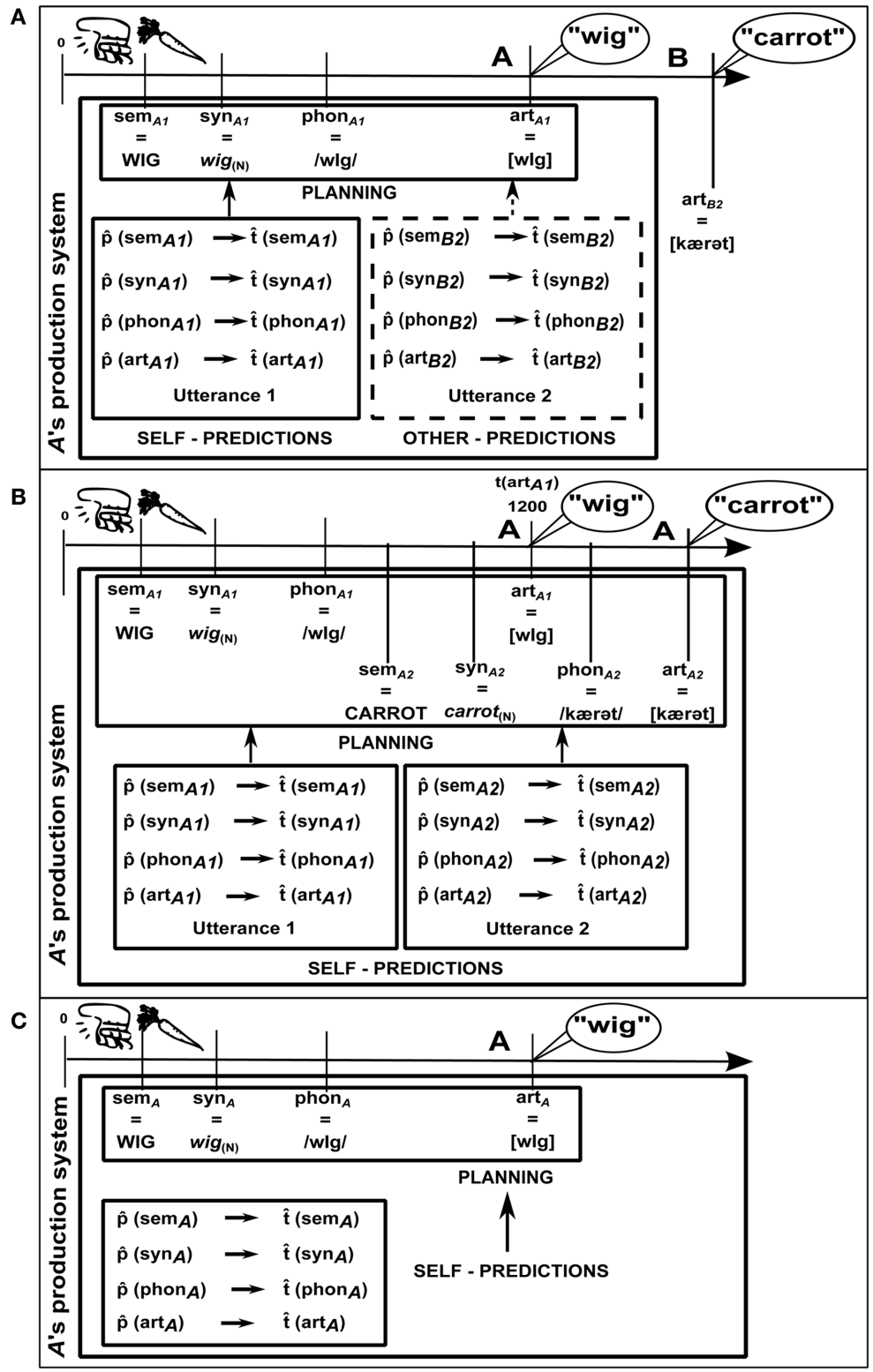


FIGURE 2 | Consecutive utterances: pictures of a wig and a carrot appear simultaneously. (A) JOINT: *A* names the left picture, then *B* names the right picture. **(B)** SOLO: *A* names the left picture, then *A* names the right picture. **(C)** NO: *A* names the left picture. Where two utterances are produced, we indicate the temporal relation between them by way of number subscripts (1 for the first utterance, 2 for the second utterance). In **(A)** art$_{B2}$ stands for the articulation stage of *B*'s utterance and p̂ (sem$_{B2}$) is the semantic content prediction that *A* generates in relation to *B*'s utterance. Time in ms. All other details as in **Figure 1**.

speech onset latencies for the first word to be affected by properties of the second word in the SOLO condition. This would reflect an influence of predictions of the second word's features on the planning of the first word. In the JOINT condition, we predict *A*'s speech onset will be similarly affected (though perhaps to a lesser degree), despite the fact that the second word is actually produced by *B*. This would show that predictions of the second word's features are computed and can affect planning of the first word also when the second word is generated by another speaker.

Additionally, the JOINT condition could be usefully contrasted to the NO condition, depicted in **Figure 2C**. The NO condition is equivalent to an instance of isolated production of a single word by *A*. Importantly, *A*'s task is the same in the NO and the JOINT conditions (i.e., producing Utterance 1), the only difference being that *B* does not produce Utterance 2 in the NO condition. The NO condition can therefore act as a control: no effect on onset latencies is expected.

Below we present various experiments that implement these ideas and discuss detailed predictions for each. Note that having the participants perform both roles is advisable, for two reasons. First, it allows data from both participants in a pair to be collected (therefore also comparisons between the behavior of the partners). Second, performing *B*'s task on half of the trials is likely to maximize the accuracy of *A*'s estimates of *B*'s timing.

Joint reversed length-effect

In Griffin's (2003) study, two pictures appeared simultaneously. The participant was told to name both pictures, avoiding pauses between the two names. She found a reversed length-effect: participants tended to initiate speech later when the first name was shorter than when it was longer; they also tended to look at the second picture more prior to speech onset and less after speech onset. Meyer et al. (2007) reported no effect on speech latencies, but they showed that the gaze–speech lag for the second picture was longer when the first name was shorter. Overall, these results seem to suggest that participants can estimate the amount of time that will be available for preparation of the second name during the articulation of the first name (Griffin, 2003).

We can therefore ask if they also estimate the time that their partner spends preparing the second name. In the SOLO condition, one participant names both pictures on a given trial; this condition is the same as Griffin (2003), except for the fact that two people are present and take turns in performing the task. In the NO condition, participants alternate in naming only the first picture, with both partners ignoring the second picture. In the critical JOINT condition, one participant names the first picture, then the other names the second picture; they alternate in performing either half of the task. We expect *B* (who has to name the second picture) to start looking at the second picture earlier (relative to when *A* starts speaking) when the first name is shorter. This would show that *B* is anticipating he will have less time to prepare his utterance when *A* is speaking. Besides, we expect *A* to initiate shorter words later than longer words. This would show that *A* is estimating *B*'s speech onset latencies and taking this estimate into account to successfully coordinate with *B* in producing a fluent utterance.

A related paradigm is based on Meyer (1996). She showed that when one participant is asked to name two pictures with a conjoined noun phrase, the auditory presentation of a distractor related in meaning to the second name delays onset latencies of the conjoined phrase. Again, if *A* contributes the first noun and *B* the second noun of the conjoined noun phrase and they have to coordinate to produce a fluent utterance (JOINT condition), we predict *A*'s speech will be affected by the relationship between the distractor and the second noun.

Joint syntactic encoding

The greater the syntactic complexity of the subject of a sentence, the longer it takes to start uttering the sentence. For example, a complex subject containing a prepositional phrase modifier or a relative clause slows down initiation times compared to a simple subject composed of two conjoined noun phrases, even when length is controlled for (Ferreira, 1991). The SOLO condition would be based on Ferreira's experiments (except for the presence of two participants): sentences could be first memorized and then produced upon presentation of a "go"-signal. In the JOINT condition, both participants would memorize the sentences. Then, depending on the cue presented at the beginning of the trial, either *A* or *B* would produce the subject (e.g., *The bike*), while their partner would contribute the rest of the sentence (e.g., *was damaged* vs. *that the cars ran over was damaged*). We expect a syntactic complexity effect on initiation times of the subject.

Active utterances are also initiated faster than the corresponding passives (Ferreira, 1994). Participants in the SOLO condition either produce sentences using a set of words provided by the experimenter or they describe pictures depicting a transitive event (e.g., of a girl hitting a boy). They are instructed to always start with the word or character presented in green (the so-called "traffic light" paradigm; Menenti et al., 2011). In this way, it is possible to control the voice of the sentence (e.g., if the boy is the first-named entity, a passive will be produced, otherwise an active). In the JOINT condition, participant *A* names only this first entity, while participant *B* produces the rest of the sentence. We expect *A*'s speech onset latencies to be slower when *B* produces a passive continuation than an active continuation; similar (or larger) results would occur in the SOLO condition, but not in the NO condition. A related paradigm could compare short vs. long continuations; it is known that more disfluencies are found at the start of longer constituents (Clark and Wasow, 1998) and it takes longer to start uttering a sentence when the subject is a conjoined noun phrase than when it is a simple noun phrase (Smith and Wheeldon, 1999).

Shared error-repair

In instances of spontaneous self-repair, people stop speaking because they detected an error in their speech and then resume with the intended output. In Hartsuiker et al. (2008), participants named pictures. On a small percentage of trials, an initial picture (the *error*) changed into a target picture (the *resumption*). Participants were told to stop speaking as fast as possible when they detected the change. In one experiment (Experiment 1), then, the same participant was asked to resume as fast as possible by naming the target picture, whereas in another experiment (Experiment 2) the task was simply to stop speaking (Hartsuiker et al., 2008).

Hartsuiker et al., 2008; see also Tydgat et al., 2011) showed that the process of stopping and the process of planning the resumption share resources: in Experiment 1, participants took longer to stop naming the error when the resumption was more difficult (through the target picture being degraded) than when it was less difficult (through the picture being intact). Moreover, there is evidence for strategic processing: when a resumption follows, people tend to withdraw resources from stopping, and instead invest them in planning the resumption while carrying on speaking. In other words, they prefer to complete the error rather than to interrupt it right away. A two-person version of Experiment 1 (stopping and resuming) would correspond to the SOLO condition, whereas a two-person version of Experiment 2 (stopping) would be our NO condition. In the critical JOINT condition, *A* stops, then *B* resumes. Therefore, *A* does not contribute the resumption. However, if she predicts that *B* will resume, we expect she will preferentially withdraw resources from stopping and complete the error, even if she does not need to invest these resources in planning the resumption.

CONCLUSION

After reviewing the literature on joint actions, we identified three mechanisms of action coordination: representational parity between self- and other-generated actions, prediction of observed actions, and integration of others' actions into the planning of one's own actions. We then claimed that similar mechanisms could underlie the coordination of utterances. We gave a comprehensive account of the type of information that could be represented about another's utterances. In considering the nature of these representations, we proposed that they are predictions generated by a forward model of one's own production system. Finally, we described two types of experimental paradigms (simultaneous productions and consecutive productions) that may prove informative as to the nature, extent, and accuracy of other-representations.

ACKNOWLEDGMENTS

Chiara Gambi is supported by a University of Edinburgh studentship. We thank Joris Van de Cavey and Uschi Cop for useful discussions.

REFERENCES

Adank, P., Hagoort, P., and Bekkering, H. (2010). Imitation improves language comprehension. *Psychol. Sci.* 21, 1903–1909.

Atmaca, S., Sebanz, N., Prinz, W., and Knoblich, G. (2008). Action co-representation: the joint SNARC effect. *Soc. Neurosci.* 3, 410–420.

Blakemore, S.-J., and Decety, J. (2001). From the perception of action to the understanding of intention. *Nat. Rev. Neurosci.* 2, 561–567.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition* 75, B13–B25.

Brass, M., Bekkering, H., Wohlschläger, A., and Prinz, W. (2000). Compatibility between observed and executed finger movements: comparing symbolic, spatial, and imitative cues. *Brain Cogn.* 44, 124–143.

Buccino, G., Binkofski, F., and Riggio, L. (2004). The mirron neuron system and action recognition. *Brain Lang.* 89, 370–376.

Caramazza, A., Costa, A., Miozzo, M., and Bi, Y. (2001). The specific-word frequency effect: implications for the representation of homophones in speech production. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1430–1450.

Clark, H. H. (1996). *Using Language.* Cambridge: Cambridge University Press.

Clark, H. H. (2002). Speaking in time. *Speech Commun.* 36, 5–13.

Clark, H. H., and Wasow, T. (1998). Repeating words in spontaneous speech. *Cogn. Psychol.* 37, 201–242.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *J. Phon.* 31, 139–148.

Cummins, F. (2009). Rhythm as entrainment: the case of synchronous speech. *J. Phon.* 37, 16–28.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385.

De Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language* 82, 515–535.

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121.

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *J. Mem. Lang.* 30, 210–233.

Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *J. Mem. Lang.* 33, 715–736.

Fowler, C. A., Brown, J. M., Sabadini, L., and Weihing, J. (2003). Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. *J. Mem. Lang.* 49, 396–413.

Garrod, S., and Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition* 27, 181–218.

Garrod, S., and Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Top. Cogn. Sci.* 1, 292–304.

Gravano, A., and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* 25, 601–634.

Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychon. Bull. Rev.* 10, 603–609.

Grosjean, F., and Hirt, C. (1996). Using prosody to predict the end of sentences in English and French: normal and brain-damaged subjects. *Lang. Cogn. Process.* 11, 107–134.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* 27, 377–442.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280–301.

Hamilton, A., Wolpert, D. M., and Frith, U. (2004). Your own action influences how you perceive another person's action. *Curr. Biol.* 14, 493–498.

Hartsuiker, R. J., Catchpole, C. M., De Jong, N. H., and Pickering, M. J. (2008). Concurrent processing of words and their replacements during speech. *Cognition* 108, 601–607.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Commun.* 53, 23–35.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–937.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., and Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286, 2526–2528.

Indefrey, P., and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition* 92, 101–144.

Keller, P. E., Knoblich, G., and Repp, B. H. (2007). Pianists duet better when they play with themselves: on the possible role of action simulation in synchronization. *Conscious. Cogn.* 16, 102–111.

Kerzel, D., and Bekkering, H. (2000). Motor activaction from visible speech: evidence from stimulus response compatibility. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 634–647.

Knoblich, G., Butterfill, S., and Sebanz, N. (2011). "Psychological research on joint action: theory and data", in *The Psychology of Learning and Motivation*, ed. B. Ross (Burlington: Academic Press), 59–101.

Knoblich, G., and Flach, R. (2001). Predicting the effects of actions: interactions of perception and action. *Psychol. Sci.* 12, 467–472.

Knoblich, G., and Jordan, G. S. (2003). Action coordination in groups and individuals: learning anticipatory control. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 1006–1016.

Knoblich, G., Seigerschmidt, E., Flach, R., and Prinz, W. (2002). Authorship effects in the prediction of handwriting strokes: evidence for action simulation during action perception. *Q. J. Exp. Psychol.* 55A, 1027–1046.

Lerner, G. H. (1991). On the syntax of sentences-in-progress. *Lang. Soc.* 20, 441–458.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci.* 22, 1–75.

Magyari, L., and De Ruiter, J. P. (2008). "Timing in conversation: the anticipation of turn endings", in *12th Workshop on the Semantics and Pragmatics of Dialogue*, eds J. Ginzburg, P. Healey and Y. Sato (London: King's college), 139–146.

Marslen-Wilson, M. (1973). Linguistic structure ans speech shadowing at very short latencies. *Nature* 244, 522–523.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696.

Menenti, L., Gierhan, S., Segaert, K., and Hagoort, P. (2011). Shared language: overlap and segregation (of) the neuronal infrastructure for speaking and listening revealed by fMRI. *Psychol. Sci.* 22, 1173–1182.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: results from picture-word interference experiments. *J. Mem. Lang.* 35, 477–496.

Meyer, A. S., Belke, E., Häcker, C., and Mortensen, L. (2007). Use of word length information in utterance planning. *J. Mem. Lang.* 57, 210–231.

Meyer, A. S., Roelofs, A., and Levelt, W. J. M. (2003). Word length effects in object naming: the role of a response criterion. *J. Mem. Lang.* 48, 131–147.

Miozzo, M., and Caramazza, A. (2003). When more is less: a counterintuitive effect of distractor frequency in the picture-word interference paradigm. *J. Exp. Psychol. Gen.* 132, 228–252.

Navarrete, E., and Costa, A. (2005). Phonological activation of ignored pictures: further evidence for a cascade model of lexical access. *J. Mem. Lang.* 53, 359–377.

Noordzij, M. L., Newman-Norlund, S. E., De Ruiter, J. P., Hagoort, P., Levinson, S. C., and Toni, I. (2009). Brain mechanisms underlying human communication. *Front. Hum. Neurosci.* 3:14. doi:10.3389/neuro.3309.3014.2009

Pezzulo, G., and Dindo, H. (2011). What should I do next? Using shared representations to solve interaction problems. *Exp. Brain Res.* 211, 613–630.

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–226.

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci. (Regul. Ed.)* 11, 105–110.

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154.

Pulvermüller, F., and Fadiga, L. (2010). Action perception: sensorimotor circuits as a cortical basis for language. *Nat. Rev. Neurosci.* 11, 351–360.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso Del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870.

Richardson, D. C., and Dale, R. (2005). Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cogn. Sci.* 29, 1045–1060.

Richardson, M. J., Marsh, K. L., and Schmidt, R. C. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 62–79.

Riley, M. A., Richardson, M. J., Shockley, K., and Ramenzoni, V. C. (2011). Interpersonal synergies. *Front. Psychol.* 2:38. doi:10.3389/fpsyg.2011.00038

Rizzolatti, G., and Craighero, L. (2004). The mirron-neuron system. *Annu. Rev. Neurosci.* 27, 169–192.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., and Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science* 326, 445–449.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Lang. Soc.* 29, 1–63.

Schriefers, H., Meyer, A. S., and Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: picture-word interference studies. *J. Mem. Lang.* 29, 86–102.

Scott, S. K., Mcgettigan, C., and Eisner, F. (2009). A little more conversation, a little less action: candidate roles for the motor cortex in speech perception. *Nat. Rev. Neurosci.* 10, 295–302.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006a). Joint action: bodies and minds moving together. *Trends Cogn. Sci. (Regul. Ed.)* 10, 70–76.

Sebanz, N., Knoblich, G., Prinz, W., and Wascher, E. (2006b). Twin peaks: an ERP study of action planning and control in coacting individuals. *J. Cogn. Neurosci.* 18, 859–870.

Sebanz, N., and Knoblich, G. (2009). Prediction in joint action: what, when, and where. *Top. Cogn. Sci.* 1, 353–367.

Sebanz, N., Knoblich, G., and Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition* 88, B11–B21.

Sebanz, N., Knoblich, G., and Prinz, W. (2005). How two share a task: corepresenting stimulus-response mappings. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1234–1246.

Shockley, K., Richardson, D. C., and Dale, R. (2009). Conversation and coordinative structures. *Top. Cogn. Sci.* 1, 305–319.

Shockley, K., Santana, M.-V., and Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 326–332.

Smith, M., and Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition* 73, 205–246.

Staub, A., and Clifton, C. J. (2006). Syntactic prediction in language comprehension: evidence from either. . .or. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 425–436.

Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14425–14430.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10587–10592.

Tian, X., and Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front. Psychol.* 1:166. doi:10.3389/fpsyg.2010.00166

Tydgat, I., Stevens, M., Hartsuiker, R. J., and Pickering, M. J. (2011). Deciding where to stop speaking. *J. Mem. Lang.* 64, 359–380.

Valdesolo, P., Ouyang, J., and Desteno, D. (2010). The rythm of joint action: synchrony promotes cooperative ability. *J. Exp. Soc. Psychol.* 46, 693–695.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467.

Vesper, C., Butterfill, S., Knoblich, G., and Sebanz, N. (2010). A minimal architecture for joint action. *Neural Netw.* 23, 998–1003.

Vissers, C. T. W. M., Chwilla, D. J., and Kolk, H. H. J. (2006). Monitoring in language perception: the effect of misspellings of words in highly constrained sentences. *Brain Res.* 1106, 150–163.

Vlainic, E., Liepelt, R., Colzato, L. S., Prinz, W., and Hommel, B. (2010). The virtual co-actor: the social Simon effect does not rely on online feedback from the other. *Front. Psychol.* 1:208. doi:10.3389/fpsyg.2010.00208

Wilkes-Gibbs, D., and Clark, H. H. (1992). Coordinating beliefs in conversation. *J. Mem. Lang.* 31, 183–194.

Wilson, M. (2001). The case for sensorimotor coding in working memory. *Psychon. Bull. Rev.* 8, 44–57.

Wilson, M., and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychol. Bull.* 131, 460–473.

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychon. Bull. Rev.* 12, 957–968.

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 593–602.

Wolpert, D. M., and Flanagan, J. R. (2001). Motor prediction. *Curr. Biol.* 11, R729–R732.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 July 2011; paper pending published: 26 August 2011; accepted: 03 October 2011; published online: 01 November 2011.

Citation: Gambi C and Pickering MJ (2011) A cognitive architecture for the coordination of utterances. Front. Psychology ***2****:275. doi: 10.3389/fpsyg.2011.00275*
This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 30 November 2011
doi: 10.3389/fpsyg.2011.00355

The dynamics of reference and shared visual attention

Rick Dale[1]*, Natasha Z. Kirkham[2] and Daniel C. Richardson[3]

[1] *Cognitive and Information Sciences, University of California Merced, Merced, CA, USA*
[2] *Centre for Brain and Cognitive Development, Birkbeck University of London, London, UK*
[3] *Cognitive, Perceptual and Brain Sciences, University College London, London, UK*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Markus Janczyk, University of Würzburg, Germany
Michael Kaschak, Florida State University, USA

***Correspondence:**
Rick Dale, Cognitive and Information Sciences, University of California Merced, Merced, CA 95343, USA.
e-mail: rdale@ucmerced.edu

In the tangram task, two participants are presented with the same set of abstract shapes portrayed in different orders. One participant must instruct the other to arrange their shapes so that the orders match. To do this, they must find a way to refer to the abstract shapes. In the current experiment, the eye movements of pairs of participants were tracked while they were engaged in a computerized version of the task. Results revealed the canonical tangram effect: participants became faster at completing the task from round 1 to round 3. Also, their eye-movements synchronized over time. Cross-recurrence analysis was used to quantify this coordination, and showed that as participants' words coalesced, their actions approximated a single coordinated system.

Keywords: language, reference, vision, attention, coordination, synchrony, interaction, communication

INTRODUCTION

> I would even say that the alterity of the other *inscribes* in this relationship that which in no case can be "posed"
> (Derrida, 1981/2004, p. 77; Translated by Bass).

To most readers, this sentence from Derrida is void of meaning. Granted it is presented without a broader context, but such words as "alterity" and "posed" are among a network of expressions that have been critiqued as lacking any clarity or substance (e.g., Putnam, 2004). Thousands of scholars carefully train to interpret these words, and use them in their own literary studies (e.g., Norris, 2002). The postmodernist vocabulary is a stark example of the process of fixing a set of shared expressions that can confuse and even frustrate those outside the clique.

This fixing process is not particular to postmodernism, however. It can be found within and across many cliques and cultures and is integral to the use and development of language. Across families and regions of England, for example, there are at least 57 words that are systematically used to refer to a television remote control, from "doofla" to "melly" (The English Project, 2008). If you do not know what "afterclap" and "manther" refer to, you can seek out an online source of modern slang. Such normative agreement can even invert the meaning of a word. "Egregious," for example, used to mean "standing out because of great virtue," but a gradual accrual of, perhaps ironic, usage has fixed its meaning as wholly negative. The fixing process can also be very rapid, taking place during the events of a single day of a small group of people with common interests.

In the present work, we aim to elucidate the behavioral microstructure of the emergence of referential vocabulary by analyzing the eye movements and computer-mouse movements of pairs of people coordinating novel expressions for unfamiliar objects. Previous studies have analyzed these emerging expressions and how long it takes for them to arise. In the current paper, we focus exclusively on what happens in the perceptuo-motor coupling dynamics between people during this emergence. Our results suggest that the gradual construction of a shared vocabulary synchronizes two people in the fine-grained dynamics of the eyes and hand.

Cognitive science has most often been in the business of studying processes of individual cognizers (Miller, 1984). But over the past 20 years the study of cognition has moved beyond individuals and into pairs or small groups of people and the environment in which they are embedded (e.g., Turvey et al., 1981; Hutchins, 1995; Clark, 1996; Hollan et al., 2000; Knoblich and Sebanz, 2006). Pairs or groups are probably, after all, the most common context of our species' behavior. Recently, detailed experimental investigation of joint activities has generated its own literature (see the collection in Galantucci and Sebanz, 2009; see also Sebanz et al., 2006). These results align with previous work arguing that groups of people in their task environment may function, in many respects, like one single cognitive system (e.g., Hutchins, 1995). One characteristic of our species that permits such fluid, multi-person functioning is our powerful communication system. People who speak the same language have a vast shared vocabulary permitting its users to help each other orient appropriately to objects in the world (e.g., see Galantucci, 2005). Whether on the hunt in the Sahara or in a restaurant with a deep menu, a shared reference scheme can organize multi-person behaviors in efficient ways.

Our results add to this view of language as a tool to organize the microstructure of cognition and action during interaction. We employed a task in which a shared reference system emerges, and examined how it transforms the behavior of those using it. Ostensibly, it permits its users to perform reference tasks much more efficiently. If you and I both know what "the jingly one" refers to, each time one of us employs it, the other can sharply orient to the appropriate referent. This skill is most often measured by completion time of these reference tasks. Here we show that something else occurs, more fundamental than simply pace of success: an emerging referential scheme induces partners in a reference task to become coupled in their visual attentional system. To show this, we focus our analysis on the eyes and hand during

a well-understood joint task used extensively in previous work: the tangram task (Krauss and Weinheimer, 1964). Previous work has studied language use and completion times in the tangram task. In our study, we do not analyze the linguistic content of the task, as it is well-understood what occurs and has been widely replicated. Instead, we go underneath those levels of analysis, and quantify the coupling between eye-movement patterns. We show that the signature of attentional coupling changes across rounds as a referential scheme is agreed upon by two task partners.

In the tangram task, pairs of participants work with a set of six unfamiliar, abstract shapes (Krauss and Weinheimer, 1964; Krauss and Glucksberg, 1969; see **Figure 1**). They see the same shapes, but arranged in a different order. One, the "matcher," must arrange her shapes to match the order of the "director." The director must use careful description in order for the matcher to succeed. Once all six shapes are re-ordered, they repeat the task. A robust pattern of change occurs as the same set of shapes are used again and again. Participants take less time to solve the task, require fewer words to do so, and end up with a jointly constructed scheme of shorthand descriptions for the shapes (Clark and Wilkes-Gibbs, 1986; see Clark, 1996, Chapter 3, for a detailed review). Once multiple rounds have been performed, the pair are capable of effectively identifying tangrams and completing the task quite rapidly. In this sense, the two people have become a coherent, functional unit (Hutchins, 1995).

The tangram task is a carefully controlled experimental context to measure this "soft-assembly" of a two-person joint system (see Shockley et al., 2009 and Marsh et al., 2009, for theoretical discussion). Because it is well known what happens at the word level in this task, here we focus exclusively on the perceptuo-motor machinery of this system[1]. We track participants in the tangram task, and analyze the eye and mouse movements across three rounds of tangram identification. Through cross-recurrence analysis, a method based on the study of coupled dynamical systems, it is possible to obtain real-time quantification of behavioral coupling as it unfolds over rounds of tangram communication (Dale and Spivey, 2005; Richardson and Dale, 2005; see Dale et al., 2011, for a comparison to other lag-based methods). These analyses show that there is extremely tight visual and motor coordination occurring in the pair, and how this coordination changes across rounds. We conclude that these properties of the tangram identification "device" are highly similar to those properties that have been identified in individual cognitive systems. With Hutchins (1995) and Sebanz et al. (2003) we argue that two-person systems exhibit the same loose-coupling under task constraints that a single cognitive processor exhibits, further demonstrating that pairs of people or beyond may serve as coherent units of analysis themselves (Tollefsen, 2002, 2006).

What does it take for two people to form "one system"? One definition, according to Hutchins (1995), is that they are *part* of a set of goals or functions that cannot be understood through any one person alone (e.g., a speed-controlling cockpit). At a finer-grained level, another way of understanding how two people come to form a functional unit is that their perceptuo-motor behavior literally takes the same shape. For example, eye movements in our task, as we show below, become more coupled from round to round, until the lag between director and matcher is not significantly different from 0 s. Their eye movements come to approximate one another. Because the tangram task is also rendering a novel referential scheme, it is both linguistic and perceptuo-motor channels that are becoming tightly aligned in order for the participants to achieve the task. In short, their various behavioral channels go from slowly achieving the task, to a loosely coupled cognitive and perceptuo-motor network: they are no longer separate individuals achieving the task, but in some sense share the same cognitive and perceptuo-motor "state space."

This outcome is not obvious given current debate in the study of discourse and psycholinguistics. Though previous work has shown a tight coupling of visual attention during dialog (Richardson et al., 2007), and has shown systematic coupling of gaze to reference (Griffin, 2001), it is unclear how this tight coupling emerges. In Richardson et al.'s (2007) work, the coupling of visual attention is based on a well-established set of words and events that interlocutors recognize and discuss (e.g., of *Simpsons* television characters). But it requires years to establish that level of expertise with language, and also requires considerable common ground. In the current study, an entrained vocabulary is assumed to emerge in just minutes, in a referential domain (tangram shapes) that is completely unfamiliar to the participants.

We thus recognized two possibilities. First, a pair may speed up in their performance as they progress through the task, but exhibit only weak and unchanging perceptuo-motor coupling characteristics. For example, the director's attention might consistently lead the matcher's all the way through each round of the task, with the maximal overlap in their eye-movements unchanging. In such a circumstance, language is speeding up only their choice performance, and not organizing their perceptuo-motor channels. A second possibility is that the two participants in this task will change flexibly together as the task unfolds, and the director and

[1]For recent investigation of speech and perceptual channels in a related problem-solving task see Kuriyama et al. (2011) and Terai et al. (2011).

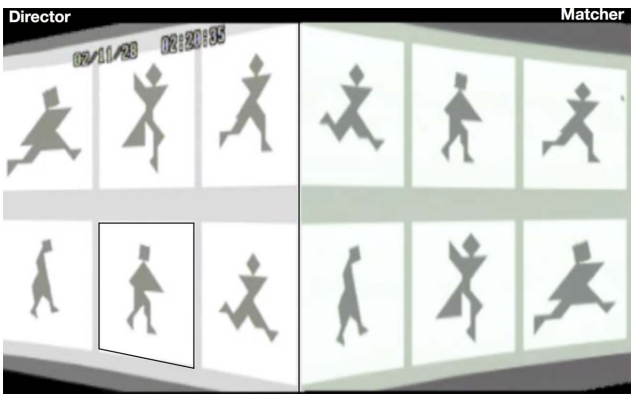


FIGURE 1 | Split screen view of an example tangram trial used in this task. The director, looking at the screen on the left, seeks a description to help the matcher select the same shape on his or her screen. Across rounds, referential language changes from detailed descriptions, such as "the guy kind of carrying the triangle," (highlighted here with a box) to simplified, entrained expressions, such as "carrying guy."

matcher come to exhibit tighter coupling dynamics. If so, the director's lead will be diminished (if not obliterated), and the two people in the task, director, and matcher, will come to have more and more locked visual attention under a referential scheme that emerges in just minutes.

EXPERIMENT

METHODS

Participants

Twenty pairs of participants were recruited from the Stanford University subject pool, and performed the tangram task for class credit. One participant in a pair was randomly assigned to the director role, and the other was assigned to matcher. Eight of these pairs did not provide mouse-movement data due to technical problems. The remaining 12 pairs formed the basis of eye-mouse analyses (see below).

Apparatus

Two eye-tracking labs on different floors of a building were used. In one of the labs an ASL 504 remote eye-tracking camera was positioned at the base of a 17″ LCD display. Participants sat unrestrained approximately 30″ from the screen. The display subtended a visual angle of approximately 26° × 19°. The camera detected pupil and corneal reflection position from the right eye, and the eye-tracking PC calculated point-of-gaze in terms of coordinates on the stimulus display. A PowerMac G4 received this information at 33 ms intervals, and controlled the stimulus presentation and collected looking time data. The second lab used the same apparatus with one difference: the display was a 48″ × 36″ back projected screen and participants sat 80″ away (this lab was designed for infants under a year old). A slightly larger visual angle of approximately 33° × 25° was subtended in this second lab. Participants communicated through the intercom feature on 2.4 GHz wireless, hands-free phones.

Stimuli

Six tangram shapes were used, similar to those used in previous work. These shapes derive from combinations of common geometric objects (squares, triangles, etc.), and many appear to be humanoid-like forms with subtle distinctions among them. These were projected in a randomized fashion in a 2 × 3 grid to both director and matcher.

Procedure

Each participant in the pair was told if s/he was a director or a matcher, and kept that roles for the duration of the experiment. They performed three rounds of the tangram task. In each, the order of the shapes was randomized for both participants. The director described each shape in turn. Whereas in the classic task, the matcher re-ordered the shapes, in our computerized version the matcher used a mouse to select the shapes in order that they appeared for the director. When the matcher identified the sixth and last shape the round ended.

Data and analysis

We extracted three behavioral signals at a sampling rate of approximately 30 Hz: (Deye) the tangram fixated by the director, (Meye) the tangram fixated by the matcher, and (Mmouse) the tangram "fixated" by the matcher's mouse cursor. For any given participant pair and communication round, three time series were thus produced: two sequences of eye movements and one sequence of mouse movements. For each round, separate analyses were conducted on the three possible alignment pairings: director's and matcher's eye movements (Deye–Meye), matcher's mouse and eye movements (Mmouse–Meye), and director's eyes/matcher's mouse (Deye–Mmouse). To explore the patterns of coordination in these pairings, we conducted a version of cross-recurrence analysis. This simply compared all time points of two time series, and generated a lag-based percentage of how much matching or "cross-recurring" (i.e., tangram fixation) took place at each lag. By plotting this percentage match, known as percentage recurrence or %REC, across all lags, we generated a *diagonal-wise recurrence lag profile* reflecting the pattern of coordination between the two time series (akin to a "categorical" cross-correlation function; see Dale et al., 2011; also see Jermann and Nuessli, 2011, for an elegant explanation).

When the %REC is largely distributed to the right or left of such a plot, it has direct bearing on the leading/following patterns of the systems producing those time series. For example, consider the top-right recurrence profile shown in **Figure 2**. This is the eye-movement %REC profile for Deye–Meye on round 1 for a particular dyad. The largest proportion of recurrent looks occurs at negative lags. This shows that at this early stage of the task, the director's eye movements are leading the matcher's (see Richardson and Dale, 2005, for more methodological detail).

Examples of time series and construction of the recurrence lag profiles are shown in **Figure 2**. To quantify how these profiles changed their position and shape across rounds, we treated the recurrence profiles as distributions of temporal data. The mean lag will be the central tendency of the overall coordination pattern, kurtosis will reflect how pointed the coordination is, and so on. Such a distribution analysis of the recurrence profile permitted us to describe quantitatively the changes in shape and position that can be seen, for example, in **Figure 2**.

For each dyad, round, and modality combination we extracted five characteristics of the recurrence lag profiles. First, we measured the overall mean recurrence across the whole profile (avg. %REC). This would be akin to measuring the mean density of a probability distribution (mean of *y*-axis values). This simply reflects, in a ±lag window, how much overall cross-recurrence is occurring between two time series. Second, we measured the maximum %REC occurring in the profile. In analysis of distributions, this is equivalent to finding the value of the maximum density (maximum *y*-axis value). This measure would reflect the maximum recurrence, achieved at one of the lags. Third, kurtosis and dispersion (SD) of the profiles were produced. The first of these measures reflects the pointedness of the coordination. A high kurtosis would indicate the presence of coordination within a small lag window, occurring for a shorter, pointed period of time; lower kurtosis would reflect a broad lag window during which states are recurrent. Dispersion (SD) has the inverse interpretation, and is calculated by treating the profile as a distribution of lags and finding the SD of the sample. Finally, we measured the central tendency (mean) of the profile. In simple distribution

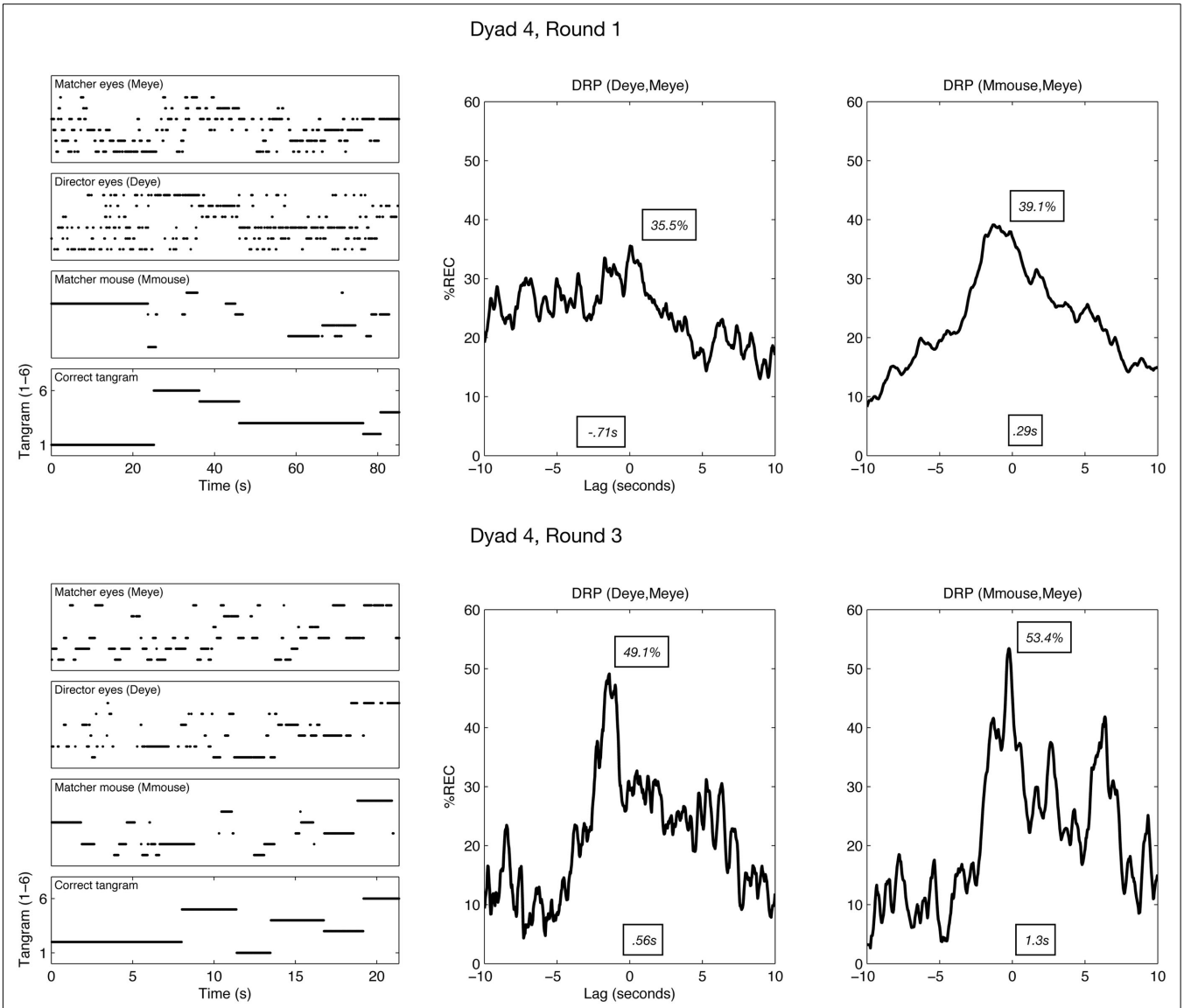


FIGURE 2 | Left column: example time series from one dyad in rounds 1 and 3. The bottom row shows, across time, the correct tangram (expressed as numeric code) that the director is attempting to get the matcher to find. Above this correct tangram is the time series for each of the analyzed channels. For example, Deye (director eyes) shows the time series of which tangram is fixated at a given moment (expressed again as a consistent code from 1 to 6). Middle/right columns: the recurrence lag profiles of pairs of these time series, with mean (s) and maximum (%) shown as examples of quantifying the profiles as a distribution (DRP = diagonal-wise recurrence profile). The profile is constructed by finding how much each time series matches (expressed as percentage recurrence, %REC) when they are lagged relative to one another. The maximum would reflect, for example, the relative point in time at which the channels are maximally aligned. See main text for more details.

analyses, this is equivalent to finding the point along the *x*-axis (here, a lag in seconds) that reflects the center of the distribution. This would measure the overall weighted center of the recurrence profile. A positive or negative mean (different from 0) would be indicative of leading or following by one of the time series (see Obtaining Distributions from Lag Profiles in Appendix for more detail).

We chose a lag window of ±10 s to explore matching between modalities. In previous work, we have found that crucial peaking of recurrence between two people is at approximately ±3 s (Richardson and Dale, 2005; Richardson et al., 2007, 2009b). We chose a wider window to ensure that our analyses both contain the key coordination region and the broader shape of the distribution.

RESULTS

Below, we first present the canonical tangram effect: participants became faster at completing the task from round 1 to round 3. Following this, we conducted a baseline analysis to show that overall coordination across the three modality pairings (Deye–Meye, Mmouse–Meye, and Deye–Mmouse) is above shuffled baseline comparisons. Finally, in a test of the profile distribution characteristics, results reveal two systems that are becoming one:

eye-movements synchronize, the matcher's eyes, and mouse are lagged relative to each other but more pointedly over rounds, and the director's eyes and the matcher's hand exhibit a distinct temporal lag. In short, the two participants, director and matcher, approximate a single coordinated system. In analyses presented below, to analyze individual distribution values across the 20 pairs, we used a linear mixed-effects model (lmer in R) treating subject as a random factor, and tangram round as the sole fixed effect. In a manner described in Baayen et al. (2008), we report *p*-values derived from Markov chain Monte Carlo (MCMC) methods calculated from *p*-values fnc in R. This analysis was chosen because it allows use of round as a continuous variable to estimate change from round to round. Where reported, approximate degrees of freedom are estimated using a Kenward–Roger correction technique described in Kenward and Roger (1997) using KRmodcomp in R (it is important to note that the MCMC significance levels are established based on simulation of the data, and *not* on the approximate degrees of freedom. These estimates are shown for convenience).

Completion time

As in previous tangram experiments (see Clark, 1996), dyads became increasingly effective at performing the task. Participants required an average of 141.5 s in the first round, 57.8 s in the second, and only 34.8 in the third. The last two rounds were significantly faster than round 1, ts > 10, ps < 0.0001. Round 3 was also carried out faster than round 2, $t(19) = 5.6$, $p < 0.0001$.

Shuffled vs. non-shuffled lag profile

We first conducted a shuffled baseline analysis for all measures. This was done by performing the same lag profile analysis but with shuffled versions of our time series, so that the temporal structure is removed. As would be expected, the total recurrence in all analyses within the ± 10 s window was substantially higher in the non-shuffled vs. shuffled conditions, ts > 7, ps < 0.0001. This main effect of shuffling held in each round when analyzed separately. In short, coordination is significant across all rounds compared to baseline, across all analyses: Deye–Meye, Deye–Mmouse, and Mmouse–Meye. The question we explore in distribution analyses below is how that coordination is organized. (Please see Which Baseline to Use? in Appendix for a discussion of use of shuffling as a reasonably conservative baseline for a data set of this size, and a comparison to other methods.)

Director–matcher eye-movement synchronization (Deye–Meye)

The recurrence lag profiles for the alignment between director's eye movements and matcher's eye movements is shown in **Figure 3A**. It revealed several significant effects across rounds. First, the overall recurrence (mean %REC) drops from round to round, $t(39) = 4.9$, $p < 0.0001$, with overall recurrence higher in round 1 (30.3%) than rounds 2 (24.5%) and 3 (21.1%; ps < 0.005). Second, there is also a main effect of round for the maximum %REC achieved, $t(39) = 2.9$, $p < 0.05$. Round 1 (39.3%) has a lower maximum %REC value than round 3 (45.0%; $p < 0.05$), with round 2 (42.1%) in between (but not significantly differing from these). It is important to note that this maximum difference may not be visible in **Figure 2**, because the maximum of the averaged profiles is not necessarily the same as the averaged of the maximum of the profiles (e.g., consider two non-overlapping normal distributions have higher average maximum, than the maximum of their average). Third, kurtosis if these distributions increases across rounds, as is indeed visible in the average profiles, $t(39) = 5.4$, $p < 0.001$. Rounds 3 (2.4) and 2 (2.1) had higher kurtosis than round 1 (1.9; ps < 0.05). Likewise, dispersion in terms of the SD (in seconds) of the profiles is decreasing from round 1 (5.5 s) to 2 (5.2 s) to 3 [4.8 s; $t(39) = 6.5$, $p < 0.001$]. Finally, the mean of this lag profile (in seconds) is changed from round to round, $t(39) = 3.0$, $p < 0.005$. The center of these profiles is shifting toward 0 s, with round 1 (-0.7 s) and round 2 (-0.8 s) significantly lower than 0 s, ts > 4, $p < 0.001$. By round 3, however, the recurrence lag profiles have an average center of 0.3 s, which is not significantly different from 0, $t(19) = 0.9$, $p = 0.4$.

Overall, the recurrence lag profiles between the eye movements of director and matcher, are becoming more sharply (higher kurtosis, lower dispersion) synchronous (center near 0) across rounds of communication. Though average %REC of the whole distribution is higher in the earlier rounds of communication, it achieves a smaller maximum, and has a distribution that is shifted away from that center of 0. By later rounds, the referential scheme synchronizes the eyes near a lag of 0 and does so without requiring long stretches of time. In short, the director and matcher are

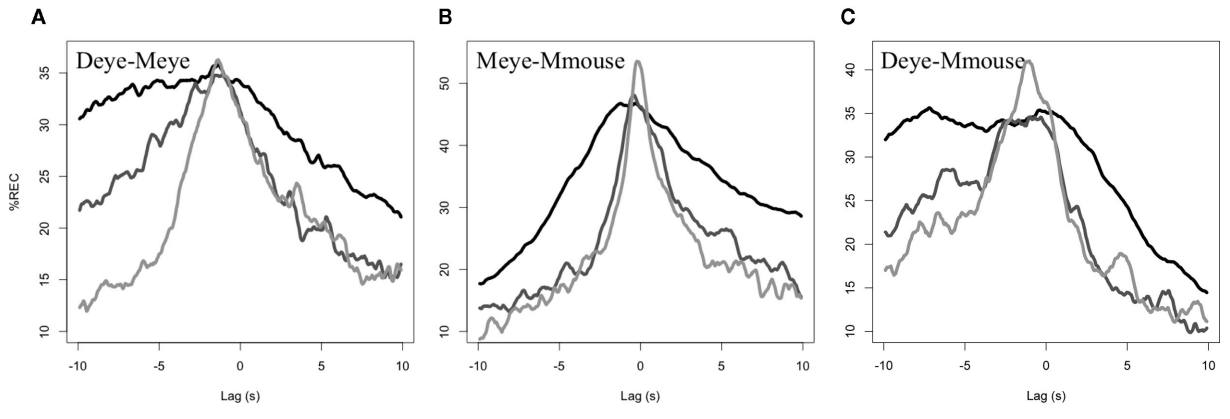


FIGURE 3 | Mean lag profiles across dyads. Round **(A)** is black, round **(B)** mid grey, and round **(C)** light grey.

coming to exhibit highly coordinated patterns of visual attention as the referential system is emerging in the task.

Matcher mouse-movement/matcher eye-movement synchronization (Mmouse–Meye)

As noted above, eight of the pairs did not supply matcher mouse tracking due to technical errors. We used the time series (Mmouse and Meye) from the remaining 12 to conduct the same linear mixed-effects analyses on the recurrence lag profile characteristics. Parallel to the statistics reported in the previous section, we obtained the following results.

Overall recurrence is again diminishing across rounds 1–3 (34, 24.7–22.3%, respectively), $t(23) = 4.2$, $p < 0.001$. Maximum recurrence is changing over rounds, with the direction of the effect exhibiting the same pattern (49.9, 52.0, and 57.9%, across rounds), $t(23) = 2.6$, $p < 0.05$. In individual comparisons, round 3 did have significantly higher recurrence than round 1 ($p < 0.05$). Kurtosis did significantly change over rounds, $t(23) = 2.6$, $p < 0.05$ (2.1, 2.4, and 2.5 from rounds 1 to 3), though dispersion did not seem to change, but is again in the same direction as seen in the previous analysis (5.1, 4.8, and 4.7 s), $t(23) = 1.6$, $p = 0.11$. The mean of the lag profile did not change, $t(23) = 0.16$, $p = 0.9$. Interestingly, however, the mean seemed highly stable from round to round (0.5,0.6,0.5 s)[2] and this mean value was significantly greater than 0, one-sample $t(35) = 4.0$, $p < 0.001$. This suggests that there is a stable leading by the eyes by approximately 520 ms overall. **Figure 3B** shows average recurrence profiles.

Though the pattern of significance is different, likely due to lessened power given lost data, the same general patterns held. The drop in average %REC and increase in kurtosis suggests that the eyes and hand are becoming more sharply coordinated in time. In addition, the stability in the mean value, and significant deviation from 0, suggests a structural limitation of the matcher's hand–eye coordination: there is consistent leading of the hand by the eye.

Direct eye-movement/matcher mouse-movement synchronization (Deye–Mmouse)

In analysis of the 12 pairs that provided Mmouse data, the following results held. First, there appears to be a drop again in mean density of %REC (29.4, 22.5, 22.1%), but this is only marginally significant, $t(23) = 1.6$, $p = 0.08$. Maximum %REC value is significantly increasing from round to round (42.9, 47.8, and 54.6%), $t(23) = 2.2$, $p < 0.05$. Kurtosis (2.1, 3.1, and 2.5) and dispersion (5.2, 4.5, and 4.6 s) did not achieve significance. Interestingly, the mean was again relatively stable in these profiles (−1.0, −1.4, and −0.9 s) indicating that the director's eyes lead the hand of the matcher by approximately 1 s, one-sample $t(35) = -3.8$, $p < 0.001$. In general, these results lack the robustness of those in Section "Director–Matcher Eye-Movement Synchronization (Deye–Meye)," but argue for an invariant of matcher's hand following the director's eyes that is perhaps predictably greater than the delay on the matcher's own eyes (see **Figure 3C** for average profiles).

[2]NB: the sign on the mean reflects the direction of leading/following by a given time series. Here, positive values indicate the matcher's eyes are leading. Negative values would have the opposite interpretation. This interpretation is simply determined by the order in which the time series are entered into analysis.

Mouse serving as spatial index?

In the previous analysis, it appears that the mouse–cursor time series maintain a kind of invariant temporal relationship with Deye and Meye – it is lagged by a certain time signature, and does not appear to change from round to round. One reason for this may be that the mouse remains stable over candidate choices, and only moves once the tangram choice has been established (e.g., clicking on the current shape it is hovering over, or moving to a new selection). This possibility is suggested in **Figure 2**, in which it can be seen that the mouse–cursor time series are relatively more stable than the eyes, and tend to remain on top of particular possible choices.

In order to test this idea quantitatively, we compared the eye-movement time series (Deye/Meye) with the matcher's mouse (Mmouse): if the mouse is serving as a kind of "holding place," then it should exhibit longer stretches of one particular event than the eyes, which are sampling the tangram visual array more freely. To do this, we measured the number of times the tangram fixated (by the eyes and "fixated" by the mouse) changes from $t - 1$ to t. We then divide this count score by the length of a given time series to obtain a percentage score for the proportion of changes occurring in the time series. When we do this, Mmouse time series change considerably less often (2.07%) than Deye (6.06%) and Meye (7.08%), $ts > 7$, $ps < 0.0001$.

One problem with this analysis, however, is that we cannot know the baseline stability of manual movements compared to eye movements under any other circumstance. It may be expected that the mouse will move less than the eyes. In order to further test the notion that the mouse is serving as a stable spatial index, we carried out an additional analysis. **Figure 4** shows trials of a given length (>15 s), averaged across all participants and trials, and plots the probability that Meye and Mmouse are on the correct tangram during the last few seconds before it is selected. The matchers' eyes

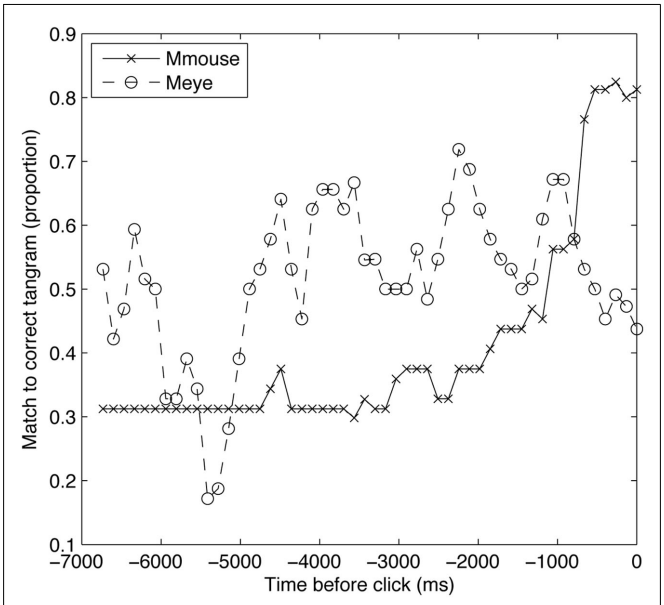


FIGURE 4 | Eye and mouse fixations on the correct tangram shape in the seconds before selection.

are more likely to be looking at the correct tangram for most of this period, as the matcher first locates the tangram and then moves the mouse to it.

Interestingly, in the last moments of the trial, Meye drops rapidly, below Mmouse. The matcher looks away from the correct tangram while their mouse remains. After listening to some of the conversations, we observed that often during the final moments of the trial, after having successfully identified a tangram, participants would look around at close competitors and confirm that they were onto the intended shape (e.g., "Ok so not the runner, the walker"). This pattern of converging upon the correct shape and then double checking other candidates can be seen in the dynamics of the eyes and hand. In particular, the use of the mouse pointer as a marker has the hallmarks of what Kirsch and Maglio (1994) called an "epistemic action": an external physical action that serves an internal cognitive function. In experiments on "spatial indexing" (Richardson and Spivey, 2000; Richardson and Kirkham, 2004) external location plays a similar role supporting cognition.

GENERAL DISCUSSION

At the beginning of the tangram task, when director and matcher have not yet become coordinated through referential expressions, the director's eyes lead the matcher's eyes. We demonstrated this through quantifying the alignment between eye movements of both people with cross-recurrence analysis. After generating a diagonal-wise recurrence lag profile, we treated it as a distribution, and quantified its characteristics. At the start of the experiment, the overall recurrence between director and matcher eye movements reflects a significant lead by the director: the profiles are shifted to the left. We asked how this coupling changes over rounds of the tangram task. This can be expressed as a test of how the profile's shape is changing, using the distribution characteristics extracted from the recurrence profile as a quantification of this change. By the final round, systematic cross-modal coordination emerged. Importantly, the recurrence profiles of director/matcher eye movements were centered at 0 s, suggesting that, on average, the director is no longer so sharply leading the matcher. It is not simply that the director and matcher achieve the task faster, but they are strongly synchronized in their perceptuo-motor activity. With the emerging interplay among multiple behavioral channels, the two participants are therefore acting as a single, coordinated "tangram recognition system." **Table 1** summarizes our basic findings.

Though the eyes synchronize, the hand's behavior may serve a separate purpose. We found in analysis of the time series that the matcher's hand remains relatively more stable than the eyes, and that it maintains a stable temporal lag relationship to the director's and matcher's eyes. The matcher's hand remains lagged, likely due to an "anchoring" to spatial indices in the visual workspace (see also Ballard et al., 1995; Brennan, 2005; Richardson et al., 2009a). As the eyes of director and matcher sample the world to be potentially responded to, the hand stays steady above candidate decisions.

This characterization of the pair as a single "system" can be understood on the backdrop of recent work on the coordination of reference domains during interaction. For example, participants in interactive tasks are subtly influenced by shared and unshared information (Richardson et al., 2007, 2009b), suggesting

Table 1 | Summary of basic findings of distribution measures across rounds.

Combo	DV	Pattern obtained across rounds (1–3)
Deye–Meye	%REC	Decreases***
	Max	Increases*
	Kurtosis	Increases***
	SD	Decreases***
	Mean	Shifts toward 0**
Mmouse–Meye	%REC	Decreases***
	Max	Increases$^{n.s.}$
	Kurtosis	Increases*
	SD	Decreases$^{n.s.}$
	Mean	No apparent change; Meye leads Mmouse by 520 ms***
Deye–Mmouse	%REC	Decreases$^{n.s.}$
	Max	Increases$^{n.s.}$
	Kurtosis	No apparent change
	SD	No apparent change
	Mean	No change; Deye leads Mmouse by 1, 113 ms***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s., not significant.

that coordination is a central component of naturalistic interactive tasks (Tanenhaus and Brown-Schmidt, 2008). Attention and comprehension are coordinated tightly as participants become accustomed to a complex referential domain (Brown-Schmidt et al., 2005, 2008). Sebanz et al. (2003) have argued that the very representations and processes used by partners in a task come to overlap simply by being co-present, and particularly by being jointly involved and aware of each other's roles during the task (see also Knoblich and Jordan, 2003; Richardson et al., 2008, 2010). Indeed, the language-as-action tradition (as described in Tanenhaus and Brown-Schmidt, 2008 and Clark, 1996), which sees one person's communication system as largely doing things to or with others, encourages a view consistent with recent perspectives on cognition as "soft-assembling" (e.g., Kugler et al., 1980) into loosely coupled functional systems during interactive tasks (Shockley et al., 2009).

The emergence of rich connections between low-level perceptual systems and high-level conceptual systems has been predicted by a number of theories (e.g., Barsalou, 1999). For example, Garrod and Pickering (2004) argue that a process of alignment cascades across all levels during interaction, and the data we present has quantified the manner in which the perceptuo-motor systems of conversants become coupled through the cascading influence of lexical entrainment (Brennan and Clark, 1996). Recent basic experimental work on individuals provides evidence that linguistic elements, such as shorthand phrases or novel labels for objects, come to organize a range of cognitive and perceptual functions, even in basic visual psychophysical tasks (e.g., Lupyan and Spivey, 2008; Huettig and Altmann, 2011). Similarly, at the level of dyads, what we have shown in the current paper is that changes in behavior during the tangram task are much deeper than a simple increase in the speed with which the task is performed. The emerging reference scheme organizes the perceptual and motor dynamics

of interlocutors. Their visual attention becomes tightly coupled, while the matcher's hand maintains an invariant temporal relationship between these two eye-movement channels – in a manner that resembles the offloading of memory during other hand–eye tasks in individuals (Ballard et al., 1995). The tight bridge between language and broader cognition is therefore a fundamental character of the fine-grained dynamics of each as they mutually influence each other during communication.

REFERENCES

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412.

Bakeman, R., Robinson, B. F., and Quera, V. (1996). Testing sequential association: estimating exact *p* values using sampled permutations. *Psychol. Methods* 1, 4–15.

Ballard, D. H., Hayhoe, M. M., and Pelz, J. B. (1995). Memory representations in natural tasks. *J. Cogn. Neurosci.* 7, 66–80.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavi. Brain Sci.* 22, 577–660.

Boker, S. M., Xu, M., Rotondo, J. L., and King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychol. Methods* 7, 338–355.

Brennan, S. E. (2005). "How conversation is shaped by visual and spoken evidence," in *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, eds J. C. Trueswell and M. K. Tanenhaus (Cambridge, MA: MIT Press), 95–129.

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493.

Brown-Schmidt, S., Campana, E., and Tanenhaus, M. K. (2005). "Real-time reference resolution by naive participants during a task-based unscripted conversation," in *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, eds J. C. Trueswell and M. K. Tanenhaus (Cambridge, MA: MIT Press), 153–171.

Brown-Schmidt, S., Gunlogson, C., and Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition* 107, 1122–1134.

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39.

Dale, R., and Spivey, M. J. (2005). "Categorical recurrence analysis of child language," in *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (Mahwah, NJ: Lawrence Erlbaum), 530–535.

Dale, R., Warlaumont, A. S., and Richardson, D. C. (2011). Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *Int. J. Bifurcat. Chaos* 21, 1153–1161.

Derrida, J. (1981/2004). *Positions*. London: Continuum International Publishing Group.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cogn. Sci.* 29, 737–767.

Galantucci, B., and Sebanz, N. (2009). Joint action: current perspectives. *Top. Cogn. Sci.* 1, 255–259.

Garrod, S., and Pickering, M. J. (2004). Why is conversation so easy? *Trends Cogn. Sci.* 8, 8–11.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition* 82, B1–B14.

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Hum. Interact.* 7, 174–196.

Huettig, F., and Altmann, G. T. M. (2011). Looking at anything that is green when hearing "frog": how object surface colour and stored object colour knowledge influence language-mediated overt attention. *Q. J. Exp. Psychol.* 64, 122–145.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cogn. Sci.* 19, 265–288.

Jermann, P., and Nuessli, M.-A. (2011). "Unravelling cross-recurrence: coupling across timescales," in *Proceedings of International Workshop on Dual Eye Tracking in CSCW (DUET 2011)*, Aarhus.

Kenward, M. G., and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53, 983–997.

Kirsh, D., and Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognit. Sci.* 18, 513–549.

Knoblich, G., and Jordan, J. S. (2003). Action coordination in groups and individuals: learning anticipatory control. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 1006.

Knoblich, G., and Sebanz, N. (2006). The social nature of perception and action. *Curr. Direct. Psychol. Sci.* 15, 99.

Krauss, R. M., and Glucksberg, S. (1969). The development of communication: competence as a function of age. *Child Dev.* 255–266.

Krauss, R. M., and Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychon. Sci.* 113–114.

Kugler, P. N., Kelso, J. A. S., and Turvey, M. T. (1980). On the concept of coordinative structures as dissipative structures: I. Theoretical lines of convergence. *Tutorials Motor Behav.* 3–47.

Kuriyama, N., Terai, A., Yasuhara, M., Tokunaga, T., Yamagishi, K., and Kusumi, T. (2011). "Gaze matching of referring expressions in collaborative problem solving," in *Proceedings of International Workshop on Dual Eye Tracking in CSCW (DUET 2011)*, Aarhus.

Lupyan, G., and Spivey, M. J. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli. *Curr. Biol.* 18, R410–R412.

Marsh, K. L., Richardson, M. J., and Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Top. Cognit. Sci.* 1, 320–339.

Miller, G. A. (1984). "Informavores," in *The Study of Information: Interdisciplinary Messages*, eds F. Machlup and U. Mansfield (New York, NY: Wiley), 111–113.

Norris, C. (2002). *Deconstruction: Theory and Practice*. New York, NY: Routledge.

Putnam, H. (2004). *Ethics without Ontology*. Cambridge, MA: Putnam.

Richardson, D. C., Altmann, G. T. M., Spivey, M. J., and Hoover, M. A. (2009a). Much ado about eye movements to nothing: a response to Ferreira et al.: taking a new look at looking at nothing. *Trends Cogn. Sci.* 13, 235–236.

Richardson, D. C., Dale, R., and Tomlinson, J. M. (2009b). Conversation, gaze coordination, and beliefs about visual context. *Cogn. Sci.* 33, 1468–1482.

Richardson, D. C., and Dale, R. (2005). Looking to understand: the coupling between speakers and listeners eye movements and its relationship to discourse comprehension. *Cogn. Sci.* 29, 1045–1060.

Richardson, D. C., Dale, R., and Kirkham, N. Z. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychol. Sci.* 18, 407–413.

Richardson, D. C., Hoover, M. A., and Ghane, A. (2008). "Joint perception: gaze and the presence of others," in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, Austin, TX, 309–314.

Richardson, D. C., and Kirkham, N. Z. (2004). Multimodal events and moving locations: eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *J. Exp. Psychol. Gen.* 133, 46–62.

Richardson, D. C., and Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition* 76, 269–295.

Richardson, D. C., Street, C. N. H., and Tan, J. (2010). "Joint perception: gaze and beliefs about social context," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Austin, TX.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn. Sci. (Regul. Ed.)* 10, 70–76.

Sebanz, N., Knoblich, G., and Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition* 88, B11–B21.

Shockley, K., Baker, A. A., Richardson, M. J., and Fowler, C. A. (2007). Articulatory constraints on interpersonal postural coordination. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 201–208.

Shockley, K., Richardson, D. C., and Dale, R. (2009). Conversation and coordinative structures. *Top. Cogn. Sci.* 1, 305–319.

Tanenhaus, M. K., and Brown-Schmidt, S. (2008). Language processing in the natural world. *Philos. Trans. R. Soc. B Biol. Sci.* 363, 1105.

Terai, A., Kuriyama, N., Yasuhara, M., Tokunaga, T., Yamagishi, K., and Kusumi, T. (2011). "Using metaphors in collaborative problem solving: an eye-movement analysis." in *Proceedings of International Workshop on Dual Eye Tracking in CSCW (DUET 2011)*, Aarhus.

The English Project. (2008). *Kitchen Table Lingo.* London: Ebury Press.

Tollefsen, D. P. (2002). Collective intentionality and the social sciences. *Philos. Soc. Sci.* 32, 25.

Tollefsen, D. P. (2006). From extended mind to collective mind. *Cogn. Syst. Res.* 7, 140–150.

Turvey, M. T., Shaw, R. E., Reed, E. S., and Mace, W. M. (1981). Ecological laws of perceiving and acting. *Cognition* 9, 237–304.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 September 2011; accepted: 10 November 2011; published online: 30 November 2011.

Citation: Dale R, Kirkham NZ and Richardson DC (2011) The dynamics of reference and shared visual attention. Front. Psychology ***2****:355. doi: 10.3389/fpsyg.2011.00355*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

APPENDIX

OBTAINING DISTRIBUTIONS FROM LAG PROFILES

Previous work has subjected cross-correlation functions to analysis (e.g., Boker et al., 2002), and the measures in this paper require a derived sample from which measures like kurtosis can be calculated. In order to treat a lag profile as a distribution, and subject it to distribution analyses, we carried out a simple translation procedure. For each time slice along the x-axis of a lag profile, we repeated that time slice's corresponding time value (e.g., in milliseconds) into a set of observations equal to some multiple (m_t) of the y-axis %REC value. In order to ensure that all lag profiles had the same sample size when subjected to distribution analyses, we used a procedure that translated the profile into $N \cong 10{,}000$ observations:

$$m_t = \text{round}(N / \Sigma_{\forall t}\ \%\text{REC}_t)$$

where $\%\text{REC}_t$ is the percentage recurrence at a give time lag t. In order to obtain the number of samples for that time value t, we simply multiply it by m_t, and the sample becomes the following collection:

$$\mathbf{x}_t = \{t, t, \ldots\} \text{ and } |\mathbf{x}_t| = \text{round}(m_t .\ \%\text{REC}_t)$$
$$\mathbf{X}_t = \cup_{\forall t} \mathbf{x}_t$$

with $\mathbf{x}_t$ as a set of observations for some time lag t, and $\mathbf{X}_t$ as the total set of observations (the union of all observations across time lags). This results in a set of observations the histogram of which resembles the original lag profile, and is composed of approximately 10,000 observations.

WHICH BASELINE TO USE?

There has been discussion of using permutation to construct baselines for these kinds of lag analyses (e.g., Bakeman et al., 1996). One recent approach is that cross-lag baselines should be assembled by "virtual pairs": Random pairs of dyads should be produced by similar analysis of time series from participants combined from separate dyads. This is important for continuous time series, for which shuffling obliterates the spectral structure of the signal (e.g., Shockley et al., 2007). However, for nominal behavior sequences of this kind, shuffling serves only to create time series the events of which occur with a probability reflecting baseline occurrence of those events (in other words, the first-order probability of looking at tangram two in a shuffled time series, at any point in time, is simply proportional to the overall frequency with which it occurs in the series).

Whether this is more or less conservative than virtual pairing, however, is not a simple question to answer. In order to test this, we developed a simple probabilistic model that produces nominal time series of the kind we analyze here. This permits large-scale exploration of the statistical impact of different baselines. We had pairs of agents ($N = 20$) take "turns" and produce 500-element nominal time series with 6 event codes (similar to the current experiment). These agents were coupled according to a simple procedure shown in **Table A1** below. The stronger the *bias* parameter, the stronger the connection between nominal sequences of agent A and B, and the greater the %REC measures.

Table A1 | Procedure for generating 2,000-element coupled symbol sequence.

Initialize agents A and B	Repeat 2,000 times: randomly choose A or B to emit symbol first with some probability (bias) make this agent reuse the symbol of the other agent from the previous turn; otherwise, choose randomly

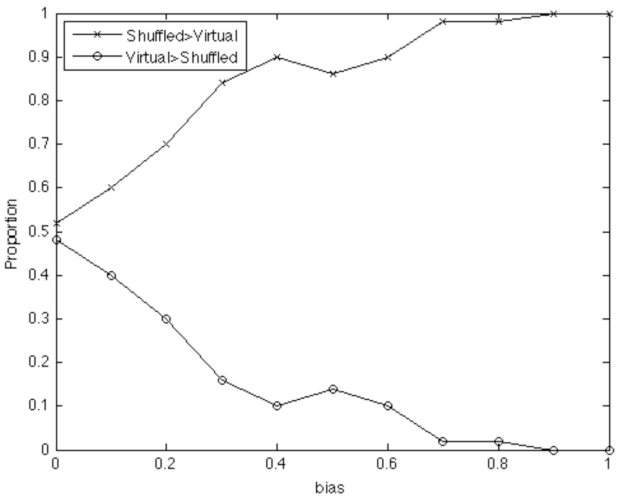


FIGURE A1 | Simple shuffling tends to produce a higher proportion of simulated baselines than the virtual pair method, especially as the 'true' coupling between systems strengthens.

We used a range of *bias* parameters, and generated 50 simulated "conversations" for each agent pair. We then did exactly the same cross-recurrence analysis over these simulations as above; we also carried out two baselines: simple shuffling and virtual pairing. The results are shown in **Figure A1** below. An average recurrence was calculated from averaging a range of ± 10 elements from the lag profile (analogous to the range ± 10 s used in the real data above, as this element range captures the coordination between agents in their lag profile). For 50 conversations (per *bias* value), the baselines were compared by assessing which would estimate a higher baseline recurrence average.

As seen in **Figure A1**, virtual pairing produces less conservative baseline scores because it estimates base-rate recurrence as *lower* than the shuffled baseline (conversely, shuffled baselines are more commonly greater in magnitude). And in fact this pattern holds the more likely there is to be an effect (i.e., with greater *bias* values, causing more tightly coupled agents). In other words, the simple shuffled baseline reflecting the base-rate probability of a particular event's occurrence provides a test that is less likely to produce a Type I error. The reason for this can be explained intuitively:

Sequences of events that hold in the original data are much less likely to overlap in virtual pairings than when shuffling occurs, because shuffling allows the individual occurrences to be distributed evenly over the time series. While the virtual pairing is more "real" in the sense that the pairs are based on the original data – the simple statistical baseline serves as a more conservative statistical basis for testing the presence of coordination. We therefore use it in this paper, as in previous papers.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 08 March 2012
doi: 10.3389/fpsyg.2012.00054

Linguistically modulated perception and cognition: the label-feedback hypothesis

Gary Lupyan*

Department of Psychology, University of Wisconsin–Madison, Madison, WI, USA

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Daniel Casasanto, Max Planck Institute for Psycholinguistics, Netherlands
David Kemmerer, Purdue University, USA

****Correspondence:***
Gary Lupyan, Department of Psychology, University of Wisconsin–Madison, Madison, WI 53706, USA.
e-mail: lupyan@wisc.edu

How does language impact cognition and perception? A growing number of studies show that language, and specifically the practice of labeling, can exert extremely rapid and pervasive effects on putatively non-verbal processes such as categorization, visual discrimination, and even simply detecting the presence of a stimulus. Progress on the empirical front, however, has not been accompanied by progress in understanding the mechanisms by which language affects these processes. One puzzle is how effects of language can be both deep, in the sense of affecting even basic visual processes, and yet vulnerable to manipulations such as verbal interference, which can sometimes nullify effects of language. In this paper, I review some of the evidence for effects of language on cognition and perception, showing that performance on tasks that have been presumed to be non-verbal is rapidly modulated by language. I argue that a clearer understanding of the relationship between language and cognition can be achieved by rejecting the distinction between verbal and non-verbal representations and by adopting a framework in which language modulates ongoing cognitive and perceptual processing in a flexible and task-dependent manner.

Keywords: language and thought, perception, categorization, labels, top-down effects, linguistic relativity, Whorf

INTRODUCTION

Are the faculties of perception, categorization, and memory – capacities humans share with other animals – shaped by the human-specific faculty of language? Does language simply allow us to communicate about our experiences, albeit with much greater flexibility compared to other animal communication systems? Or, does language also transform cognition and perception, allowing humans to access and manipulate mental representations in novel ways? This question has been of longstanding interest to philosophers (see Lee, 1996 for a historical review), and goes to the core of understanding human cognition (Carruthers, 2002; Spelke, 2003). Many have speculated on the transformative power of language on cognition (James, 1890; Whorf, 1956; Cassirer, 1962; Vygotsky, 1962; Dennett, 1994; Clark, 1998). A growing number of studies show that language can exert rapid and pervasive effects on putatively non-verbal processes. For contemporary reviews of the "language and thought debate" (see Gumperz and Levinson, 1996; Gentner and Goldin-Meadow, 2003; Gleitman and Papafragou, 2005; Casasanto, 2008; Boroditsky, 2010; Wolff and Holmes, 2011). Despite progress on the empirical front showing apparent effects of language in domains ranging from basic perceptual tasks such as color perception (see below), motion perception (Meteyard et al., 2007), visual search (Lupyan, 2008a), and simple visual detection (Lupyan and Spivey, 2010a), to categorization in infancy (e.g., Waxman and Markow, 1995) and adulthood (Lupyan et al., 2007), to recognition memory (e.g., Lupyan, 2008b; Fausey and Boroditsky, 2011) and relational thinking (Loewenstein and Gentner, 2005), there has been a lack in progress on the theoretical front. In this work, I will argue that significant theoretical progress can be made by taking a interactive-processing perspective (e.g., McClelland and Rumelhart, 1981) on the question of the relationship between language and thought.

The paper is divided into four parts: First, I discuss an apparent paradox that has stymied both critics and proponents of the "language and thought" research program (Gleitman and Papafragou, 2005; Wolff and Holmes, 2011): how can effects of labels be both deep, apparently affecting basic even perceptual processing, and yet be easily disrupted by manipulations such as verbal interference? Second, I present a proposed solution to the paradox in the form of the *label-feedback hypothesis*, on which the classic distinction between verbal and non-verbal processes is replaced with an emphasis on the role of language as a *modulator* of a distributed and interactive system (see also Kemmerer, 2010). Third, I review some empirical data from the domains of visual perception, categorization, and memory, that are difficult to reconcile with common assumptions in contemporary literature on language and thought, but are naturally accommodated by the label-feedback hypothesis. Finally, I briefly discuss the implications of taking an interactive-processing on the question of linguistic relativity.

THE FRAGILITY OF LINGUISTIC EFFECTS ON COGNITION AND PERCEPTION: A PARADOX?

One domain that has received a considerable amount of attention in the language and thought literature is that of putative effects of language on color categorization and color perception. Shortly after the posthumous publication of Benjamin Lee Whorf's essays (Whorf, 1956), the philosopher Max Black published a critique in which he commented on Whorf's now-famous passage: "We dissect nature along lines laid down by our native languages. Language is not simply a reporting device for experience but a

defining framework for it." (p. 213). Black remarked that Whorf's word-choice engendered confusion:

> "To dissect a frog is to destroy it, but talk about the rainbow leaves it unchanged. The case would be different if it could be shown that color vocabularies influence the perception of colors, but where is the evidence for that?" (Black, 1959, p. 231).

There is now a large and rapidly increasing number of findings showing just such effects: cross-linguistic differences in color vocabularies can cause differences in color categorization with concomitant effects on color memory and, indeed, color perception (Davies and Corbett, 1998; Davidoff et al., 1999; Roberson et al., 2005, 2008; Daoutis et al., 2006; Winawer et al., 2007; Thierry et al., 2009). For example, Winawer et al. (2007) presented English and Russian speakers with color swatches showing different shades of blue. Russian, unlike English, lexicalizes the category blue with two basic-level terms: "siniy" for darker blues and "goluboy" for lighter blues[1]. The subjects were asked to perform a simultaneous *xAB* task, deciding as quickly as possible whether a top color (*x*) exactly matched a color on its left (*A*) or on its right (*B*). The categorical relationship between the color *x* and the non-matching color was varied such that, for Russian speakers, the two colors were sometimes in the same lexical category and sometimes in different categories. All colors were in the "blue" category for English speakers. The results showed a categorical perception effect for Russian speakers only, as evidenced by slower reaction times (RTs) on within-category than between-category trials.

A possible mechanism by which cross-linguistic differences in categorical color perception can be produced is gradual perceptual warping caused by learning. On this account, long-term experience categorizing the color spectrum using language gradually warps the perceptual representations of color resulting in more similar representations of colors in the same category (i.e., those labeled by a common term) and/or less similar representations of colors grouped into distinct categories (i.e., those labeled by distinct terms). That is, learning and using words such as "siniy" and "goluboy" provides categorization practice that results in the gradual representational separation of the parts of the color spectrum to which the labels are applied. Different labeling patterns (using the generic term "blue") are therefore predicted to produce different patterns of discrimination across the color spectrum. This standard account of learned categorical perception (Goldstone, 1994, 1998; Goldstone and Barsalou, 1998) has been applied to the color domain, and as predicted, training individuals on a new color boundary can induce categorical perception (Ozgen and Davies, 2002).

On the perceptual learning account, once labels have provided sufficient categorization training for perceptual warping to occur, the warped perceptual space remains. And yet, a growing number of studies show that when participants are placed under conditions of verbal interference that is presumed to decrease the on-line influence of language, cross-linguistic differences seem to disappear. For example, Winawer et al. (2007) found that when Russian-speaking subjects were placed under verbal interference, within-category comparisons no longer took longer than between-category comparisons[2] (see also Roberson and Davidoff, 2000; Pilling et al., 2003; Gilbert et al., 2006; Drivonikou et al., 2007; Wiggett and Davies, 2008; cf. Witzel and Gegenfurtner, 2011). This bleaching effect of verbal interference is seen in other domains as well. For example, English and Indonesian-speaking monolinguals show memory patterns consistent with their language: better memory for different tenses in English than Indonesian, which does not require morphological tense markers (Boroditsky, 2003). The difference in memory between Indonesian and English speakers was attenuated with verbal interference.

Further evidence of the transient nature of effects of language on cognition comes from studies of the consequences of language impairments on putatively non-verbal processes (putative in the sense that if some cognitive process can be shown to be affected by language, is that process still non-verbal?). The logic as articulated by Goldstein (1924/1948) is that if language is involved in not only communicating thoughts but somehow "fixating" them, then language impairments should produce cognitive impairments. Indeed, as noted by Goldstein (see Noppeney and Wallesch, 2000 for review), individuals with aphasia appear to have a number of deficits that appear on their surface to have little to do with language. A particular difficulty is posed by categorization tasks requiring grouping on a particular dimension. In an effort to further distil this deficit, Cohen and colleagues concluded that ". . .aphasics have a defect in the analytical isolation of single features of concepts" (Cohen et al., 1980, 1981), yet are equal to control subjects "when judgment can be based on global comparison" (Cohen et al., 1980). In their examination of the anomic patient LEW, Davidoff and Roberson reached a similar conclusion, arguing that when a grouping task requires attention to one category while abstracting over others, LEW is "without names to assist the categorical solution." (Davidoff and Roberson, 2004, p. 166). In a recent study designed to examine the categorization-aphasia link more exhaustively, Lupyan and Mirman (under review) found that a group of patients with aphasia (selected on the basis of having varying levels of naming impairments) were specifically impaired on a categorization task requiring focusing on a specific dimension, e.g., selecting all the pictures of red objects from color images of familiar objects. The patients were selectively impaired on trials requiring categorizing by specific isolated dimensions, but had performance similar to controls on trials which required more global categorization such selecting objects typically found in a laundry room. Critically, the patients' impairment on this non-verbal task was best predicted by their performance on a standard confrontation naming test (PNT; Roach et al., 1996). Naming performance continued to predict categorization performance controlling for semantic impairments and general location

[1]There has been some confusion regarding the primacy of color terms such as "navy" in English. The crucial cross-linguistic difference here lies not so much in the frequency, ambiguity, or accessibility of the term "siniy" in the minds of Russian speakers versus the term "navy" in the minds of English speakers. Rather, the difference lies in the presence of a generic term "blue" in English and the lack of such a term in Russian. An inverse situation occurs in the domain of body part terms: The Russian word "ruka" (arm including the hand) has no corresponding generic term in English.

[2]Verbal interference actually *reversed* the usual categorical perception effect with within-category matching now taking less time than between-category matching (see also Gilbert et al., 2006 for a similar reversal). This odd pattern of results awaits an explanation.

of the lesion. These data do not suggest that successful categorization *depends* on an intact naming abilities, but that the two are intertwined such that naming impairments contribute to categorization impairments, particularly when the task requires isolating specific dimensions and cannot be accomplished through overall similarity[3].

Convergent evidence for the interactive relationship between language and categorization comes from a study in which I used verbal interference to attempt to simulate some of the categorization impairments that have been previously reported to be concomitant with naming impairments. Lupyan (2009) tested college undergraduates on an odd-one-out task in which participants were presented with triads of pictures or words and had to select the one that did not belong on some specific criterion, such as real-world size. On other trials, the task required selecting a picture or word that did not belong based on more thematic or functional relationship.

When tested with this task, the anomic patient LEW was selectively impaired in making size and color, but not function/thematic judgments (Experiment 7, Davidoff and Roberson, 2004). Healthy subjects undergoing verbal (but not visual) interference of the same type as used to bleach effects of language on color perception, showed a performance profile very similar to that of the anomic patient LEW.

THE PARADOX DISTILLED

The paradox then is this: if effects of language on perceptual processing are "Whorfian" in the sense of changing the underlying perceptual space (i.e., warping perception), then how can the space be "unwarped" so easily? Similarly, if language affects categorization by providing additional training opportunities, why would language impairments produce categorization impairments? In a recent debate hosted by *The Economist* on the proposition "The language we speak shapes how we think," Lila Gleitman remarked on the interpretation of the types of effects of language on color discussed above with the following observation:

> . . .here is the usual finding: "Disrupting people's ability to use language while they are making colour judgments eliminates the cross-linguistic differences." What is puzzling is why [Boroditsky] thinks this is a "pro" argument. In fact, it is the "con" argument, namely that the underlying structure and content of "thought" and "perception" are unaltered by palpable and general differences in language encoding (Gleitman, 2010).

This argument in one form or another has been invoked by a number of critics (Gleitman and Papafragou, 2005; Dessalegn and Landau, 2008; Li et al., 2009). The reasoning seems to be that if linguistic influences on categorization and perception can be removed so easily (or conversely, appear after only a brief training period, e.g., Boroditsky, 2001; cf. January and Kako, 2007), then they must be superficial. Put another way, according to this critique, if an influence of language on, for example, color perception can be disrupted via a verbal manipulation, does this not mean that language was affecting a verbal process all along and therefore the effect is of language on language rather than language on perception? This rationale appears to rest on two assumptions: First, language is assumed to be a medium (a "transparent medium" even, H. Gleitman et al., 2004, p. 363). On this view, words *map onto* concepts, which are, by definition, independent of words (e.g., Gopnik, 2001; Snedeker and Gleitman, 2004; Gleitman and Papafragou, 2005). The second assumption is of a strict separation between verbal and non-verbal processing, and consequently between verbal and non-verbal representations. (This assumption is also evident in the "thinking for speaking" framework articulated by Slobin, 1996). Accepting these two assumptions, it is indeed puzzling how the sorts of effects of language on color categorization and perception discussed above can be simultaneously pervasive and fragile: if language alters concepts, should not these altered concepts persist regardless of how language is deployed on-line?

The *label-feedback hypothesis* is an attempt to reconcile this apparent paradox of how effects of language can be so vulnerable to interference while at the same time exerting apparently pervasive influence on basic perceptual processing (e.g., see Liu et al., 2009; Thierry et al., 2009; Mo et al., 2011 for effects of language on early visual processing in the domain of color perception). As I will argue, the reason these effects are sensitive to manipulations such as verbal interference is that many language exerts effects on perception by modulating ongoing perceptual processing on-line. This modulation, insofar as it is rapid and automatic, constitutes a change in the functional structure and content of "thought" referred to by Gleitman because language and thought are part of a distributed interactive system. As articulated by Whorf himself:

> Any activations [of the] processes and linkages [which constitute] the structure of a particular language. . . once incorporated into the brain [are] all linguistic patterning operations, and all entitled to be called thinking (Whorf, 1937, pp. 57–58 cited in Lee, 1996, p. 54).

A note of caution is in order: Viewing language as a part of an inherently interactive system with the capacity to augment processing in a range of non-linguistic tasks does not mean that performance on every task or representations of every concept are under linguistic control. Rather, the argument is that learning and using a system as ubiquitous as language has the potential to affect performance on a very wide range of tasks. A fruitful research strategy may be therefore to investigate what classes of seemingly non-verbal tasks are influenced by language (and which are not), and on what classes of tasks cross-linguistic differences yield consistent differences in performance. This point is expanded below in the Section "Implication of the Label-Feedback Hypothesis for 'Language and Thought' Research Program."

[3]Kemmerer et al. (2010) tested a large and very diverse group of brain-damaged patients on a battery of tasks including naming, word–picture matching, and attribute selection (e.g., deciding which picture depicts an action that is most tiring). The deficit profile was a complex one with patients showing virtually every pattern of dissociation between the tasks. Interestingly, naming performance was significantly correlated with performance on the picture-attribute task, but not at all with the picture-comparison task. It remains to be determined if these patterns of association reflect differences in the degree to which the tasks require selection of specific dimensions versus reliance on global association (Lupyan, 2009; Lupyan et al., under review; see also Sloutsky, 2010).

FROM PERCEPTION TO CATEGORIZATION TO VERBAL LABELS AND BACK AGAIN: THE LABEL-FEEDBACK HYPOTHESIS

Perceiving a stimulus as meaningful depends on (perhaps even requires) representing the stimulus in terms of a larger class. Consider that even a task as simple as deciding whether two "identical" objects, presented simultaneously in different locations are the "same" requires the observer to ignore that they are different by virtue of their positions. In the short-story *Funes the Memorius*, Borges describes a man incapable of categorization:

> "It was not only difficult for him to understand that the generic term dog embraced so many unlike specimens of differing sizes and different forms; he was disturbed by the fact that a dog at three-fourteen (seen in profile) should have the same name as the dog at three-fifteen (seen from the front)" (Borges, 1942/1999, p. 136).

Naming both of the above instances of dogs as a "dog" requires representing both as members of the same class – one which is associated with the label "dog." Clearly, naming depends on categorization. But does language, and the act of naming in particular, play an active role in the categorization process itself? In this section, I argue that names (verbal labels) play an active role in perception and categorization by selectively activating perceptual features that are diagnostic of the category being labeled. Critically, although this top-down augmentation of perceptual representations by language is likely to be in play, to some degree, even during passive vision, it can be up- or down-regulated through linguistic manipulations such as brief verbal training/verbal priming and verbal interference.

On the present view, categorization is the *process* by which detectably different (i.e., non-identical) stimuli come to be represented as identical, in some respect (see Lupyan et al., under review for discussion). Categorizing a stimulus thus involves changing its representation. However, placing two objects into the same category does not, logically, imply a change to their perceptual representations which on some accounts are impenetrable to the influence of conceptual categories (e.g., see Pylyshyn, 1999; Macpherson, 2012). In groundbreaking work, Goldstone and colleagues (Goldstone, 1994; Goldstone and Hendrickson, 2010 for review) showed that the categorization process alters perception itself. In a typical study, participants were trained to respond to items that parametrically vary on one or more dimensions with some belonging to "Category A" and others to "Category B" (Goldstone, 1994), or to discriminate between individuals belonging to a "club" and those not belonging (Goldstone et al., 2003). Following this training, visual discrimination ability is assessed (while controlling for effects of categorization from those of mere exposure[4]) and compared to visual discrimination prior to training or to discrimination following a control training task. A significant change in the perception of dimensions relevant to the categorization task suggests that categorization experience altered the visual appearance of the items being categorized. Rather than just being *mediated* by the category responses (i.e., participants judging two stimuli as more similar by virtue of their belonging to the same category), the experience of categorization was found to warp perception, sensitizing some regions of perceptual space (e.g., those close to the category boundary; Goldstone, 1994). This warping effect affected the relationship between trained and novel stimuli–an effect argued by the authors to be incompatible with an effect of categorization on the decision process only (Goldstone et al., 2001).

Goldstone and colleagues' work on perceptual warping and learned categorical perception (e.g., Goldstone et al., 2001) provides a potential mechanism by which language may augment categorization. Because each act of naming is an act of categorization, learning to label some colors "green" and others "blue," provide a type of category-training which, over time, is expected to help pull apart the representations and resulting in decreased representational overlap between the two classes of stimuli. But how can one reconcile the perceptual warping process with the fragility of language-modulated effects outlined above?

The *label-feedback hypothesis* proposes that language produces *transient* modulation of ongoing perceptual (and higher-level) processing. In the case of color, this means that after learning that certain colors are called "green," the perceptual representations activated by a green-colored object become warped by top-down feedback as the verbal label "green" is co-activated. This results in a temporary warping of the perceptual space with greens pushed closer together and/or greens being dragged further from non-greens. Viewing a green object becomes a hybrid visuo-linguistic experience. Knowing that some colors are called green means that our everyday experiences of seeing become affected by the verbal term, which in turn makes the visual representation more categorical. This modulation can be increased – up-regulated – by activating the label to a greater than normal degree as when a participant hears a verbal label prior to seeing a visual display. Conversely, verbal interference is one way to down-regulate the activation of labels leading to reduced influences effect of language on "non-verbal" processing.

To illustrate how language can affect perceptual representations, consider a task in which subjects view briefly presented displays of the numerals 2 and 5, with several from each category presented simultaneously. The task is to attend to just the 5s and to press a button as soon as a small dot appears around one of the numerals. The more selectively participants can attend to the 5s, and just the 5s, the better they ought to perform. Before some trials, participants actually hear the word "five." This cue constitutes entirely redundant information because participants already know what they should do on each trial. The task of attending to the 5s remains constant for the entire 45-min experiment, thus the word "five" tells them nothing they do not already know. Yet, on the randomly intermixed trials on which they actually hear the word, participants respond more quickly (and, depending on the task, more accurately; Lupyan and Spivey, 2010b). This type of facilitation occurs even when the items are seen for only 100 ms, a time too brief to permit eye movements. Similar effects are obtained with more complex items such as pictures of chairs and tables. The linguistic facilitation is also transient. If too much time is allowed to elapse between the label and the onset of the display (more than ∼1600 ms. in this case), no facilitation is seen. In fact, obtaining

[4] See Folstein et al. (2010) for a recent study of the role of mere exposure to exemplars on subsequent category learning.

such effects is only possible if hearing a word has a *transient* effect on visual processing; if the facilitation due to hearing a word carried through the entire experiment, the difference between the intermixed label and no-label trials would quickly vanish. Yet the difference persisted, in most cases through the entire experiment lasting for hundreds of trials (Lupyan and Spivey, 2010b) which was only possible if hearing a label affected perceptual processing in a transient, on-line manner.

According to the label-feedback hypothesis, hearing the word "five" activates visual features corresponding to 5s, transiently moving the representations of 5s and 2s further apart, while making the perceptual representations of the various 5s on the screen more similar, and thereby easier to simultaneously attend. Notice that this task did not *require* identification or naming. Verbal labels were certainly not needed to see that 2s and 5s are perceptually different. Yet, overt language use – a hypothesized "up-regulation" of the linguistic modulation normally takes place during perception – had robust effects on perceptual processing.

In other studies, my colleagues and I have shown that hearing similarly redundant words can improve performance in a pop-out visual search (Lupyan, 2008a) and improves search efficiency in more difficult search tasks (Lupyan, 2007). Hearing a label can even make an invisible object visible. Lupyan and Spivey (2010a) showed that hearing a spoken label increased visual sensitivity (i.e., increased the d') in a simple object detection task: simply hearing a label enabled participants to detect the presence of briefly presented masked objects which were otherwise invisible (see also Ward and Lupyan, 2011 who showed that hearing labels can make visible stimuli suppressed through continuous flash suppression).

A SIMPLE MODEL OF ON-LINE LINGUISTIC EFFECTS ON PERCEPTUAL REPRESENTATIONS

A simple model implementing the idea of labels as modulators of lower-level representations is shown in **Figure 1**. The model is implemented as a fully recurrent neural network (Rumelhart et al., 1986). Solid lines denote feedforward connections and dashed lines denote feedback connections. In this implementation, the perceptual layer is provided with a feature-based input of a current object. The model is trained on two categories instantiated as a distortion from one of two category prototypes (for a more detailed description, see Lupyan, in press). Let us arbitrarily call one category "chairs" and the other "tables." During training, the model learns to produce names, e.g., to produce the label "chair" given one of the chairs, and comprehend names: given the label "chair," it activates properties characteristic of chairs. Due to the one-to-many mapping between-category labels and category exemplars the network cannot know which particular object is being referred to when presented with just the category label. It is this one-to-many mapping that allows the network to generalize and make inferences to un-seen properties. Because some properties (e.g., having a back) are more closely correlated with category membership than other properties (e.g., being brown) the category labels become more strongly associated with properties that are typical or diagnostic of the denoted category, and dissociated from properties that are not diagnostic of the category.

Following this training, we can examine what happens to representations of category exemplars when the label is allowed to feed back on the activity in the perceptual layers. **Figure 2** shows a principal-components analysis (PCA) of the perceptual representations of exemplars from two categories learned in the context of labels. In **Figure 2A**, the label is endogenous to the network. The network produces the label itself in response to the perceptual input, and the label is then allowed to feedback to affect the visual representations. This corresponds to what is hypothesized to occur in the default case: perceptual representations are modulated on-line by verbal labels via top-down feedback. In **Figure 2B** the labels are prevented from affecting the representations on-line by disabling the name-to-hidden-layer connections. The category separation observed in this PCA plot is due entirely from bottom-up perceptual differences between the two categories. This situation is logically equivalent to a verbal interference condition (although in reality, label activations are only one kind of top-down influences affecting visual processing). In **Figure 2C** the labels are provided exogenously to the network along with the perceptual input. This case is equivalent to the label trials in the experiment described above (Lupyan and Spivey, 2010b). Much clearer category separation is observed. Insofar as correct categorization depends on representing similarities between exemplars, it is facilitated by the influence of labels. There is a cost to this enhanced categorization. The more categorical representations produced by the labels are beneficial for categorization-type tasks, but reduce accuracy in the representation of the idiosyncratic properties of individual exemplars. Indeed, when participants are shown pictures of chairs and tables and are asked to label some of them with the category labels ("chair" and "table"), they show poorer subsequent recognition of items that they labeled (Lupyan, 2008b).

The simple model shown in **Figure 1** can be extended to help understand how label-feedback may affect performance in categorization tasks such as those requiring the isolation of specific dimensions – impaired in aphasia and under conditions of verbal interference. Feedback from the activation of a dimensional label such as "size" or "color" is predicted to have the same kind of cohering effect – facilitating the grouping of objects by their dimensions. This role of labels in realigning representations is one way to explain the facilitatory effect of labels in relational reasoning (Kotovsky and Gentner, 1996; Ratterman and Gentner, 1998; Gentner and Loewenstein, 2002).

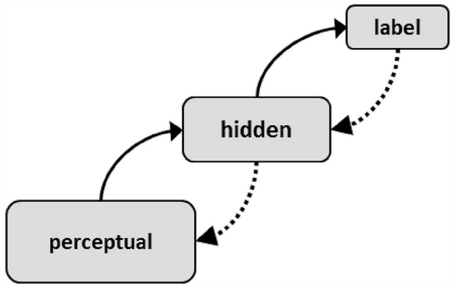


FIGURE 1 | A schematic of a neural network architecture for exploring on-line effects of labels on perceptual representations. See text for description.

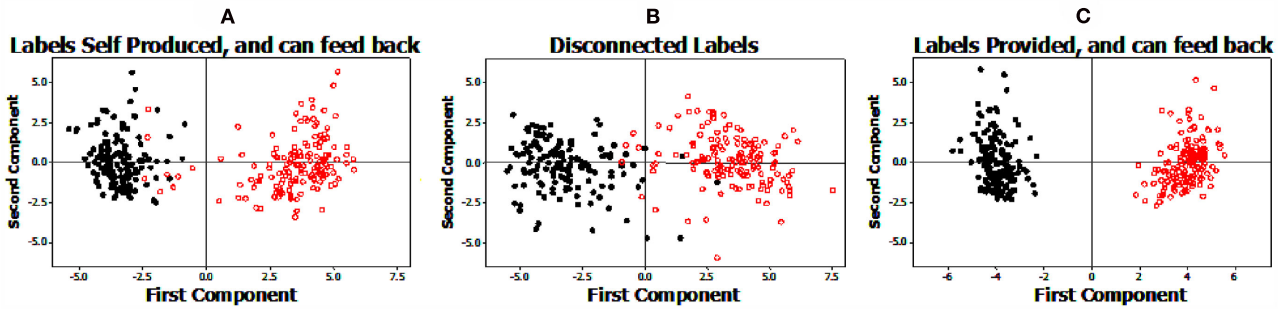


FIGURE 2 | Principal-components analyses from a connectionist simulation showing the influence of category labels on the perceptual representations. The simulation uses the network architecture shown in **Figure 1**. Each dot represents an item from one of two categories, denoted by separate colors and its location. Category structure is enhanced when labels, activated by the network, are allowed to feed back onto the perceptual layer **(A)**. When this feedback of labels is disrupted by blocking the flow of activity from the label to the hidden layer, representations revert to reflecting the perceptual structure of the stimulus space **(B)**. Categorization is further enhanced when the label is provided to the network exogenously **(C)**. See text for additional details.

ON-LINE VERSUS SUSTAINED EFFECTS OF LABELS ON PERCEPTION AND COGNITION

The demonstrations of the effects of labels on perceptual processes discussed above focused on transient effects such as those produced by overtly hearing a category name. Finding that language influences visual processing, but only in the few seconds immediately after we hear a word, while curious, is clearly of limited theoretical import. The key assumption in such experiments (e.g., Lupyan, 2007, 2008a; Lupyan and Spivey, 2010a,b) is that overt presentation of labels (or, as shown by Lupyan and Swingley, in press, language production in the form of self-directed speech) can exaggerate what is hypothesized to be the normal on-line influence of language on task performance. Verbal interference, on this view, is a comparable *down*-regulation of language. Such manipulations can shed light on the "normal" function played by language in cognition and perception. In this section I briefly review some findings suggesting that perceptual processes are influenced rapidly and automatically by language. That is, the normal state in adults is closer to **Figure 2A** in which automatically activated labels modulated perceptual representations, than **Figure 2B** in which perceptual representations mapped onto category labels, but were impermeable to linguistic feedback.

Consider a task in which an observer is presented with two stimuli and needs to determine, as quickly as possible, whether they are visually identical. Naturally, the more subtle the differences, the more difficult the judgment. Consider now the letter pairs B-b and B-p. The letters in each pair are visually equidistant, but *conceptually* B-b are more similar than B-p. Despite this conceptual difference, reaction times (RTs) for B-p and B-b judgments are equivalent when the two letters are presented simultaneously (Lupyan, 2008a). However, when the second letter is presented $\geq$150 ms. after the first (with the first still present on the screen), B-b judgments become more difficult than B-p judgments (Lupyan et al., 2010b). We claimed this occurs because during this delay, the representation of the first letter becomes augmented by its conceptual category, rendering "B" more similar to "b" and more distinct from "p." This effect is further enhanced when subjects actually *hear* the letter name (Lupyan, 2008a), i.e., up-regulating language appears to exaggerate the categorical perception effect.

Although these results show basic perception to be dynamically influenced by conceptual categories, the results do not directly address the role played specifically by the category names. This question is beginning to be addressed using the work described below.

Using fMRI, Tan et al. (2008) showed that in a same-different color discrimination task, similar to the simultaneous condition of the B-p task described above, Wernicke's area (posterior part of BA 22) showed greater activity for easy-to-name versus hard-to-name colors suggesting its automatic activation in this non-verbal task. Although the authors attempted to interpret the selective activity in terms of the effects of language on visual discrimination, clearly, no such causal attribution of the neural activity can be made; its activity may be consistent with activation of color names, but does not indicate that this activity affects visual processing. On the current account, such causal effects are exactly what is expected, with category effects in vision emerging (in some part) due to activation of category *names*. One way to test this prediction is by disrupting the activity and measuring its outcome. In a recent study, we administered TMS to Wernicke's area while participants performed the B-p/B-b same-different task (Lupyan et al., in preparation). Insofar as slower responses to B-b relative to B-p are the result of label-feedback, disrupting this activity should eliminate the RT difference between B-p and B-b stimuli. The results showed that an inhibitory stimulation regime completely eliminated the RT difference between responding "different" to B-p and B-b letter pairs. Control stimulation to the vertex had no effect. To my knowledge, no theory of visual processing classifies Wernicke's area (posterior superior temporal gyrus) as "visual." That disruption of activity in this region alters behavioral responses on a visual task supports the hypothesis that the effects of conceptual categories (here, letter categories) on visual processing are subserved in part by a classic language area, stimulation of which possibly disrupts its usual modulation of neighboring posterior regions of the ventral visual pathway.

The transient effects of labels on perception described above may be special cases of normally occurring top-down modulations of vision by linguistic, contextual and other "cognitive" factors. An example of such modulations of a more sustained nature can be

seen when one examines the role of meaningfulness in vision. As might be expected, it is easier to recognize and discriminate meaningful entities than meaningless ones. For example, it takes about 200 ms. longer to recognize that the items in the pair ꟼ/ꟼ are physically identical than it does to make the same judgment for ꟼ/ꟼ or b/b (Lupyan, 2008a). The stimuli ꟼ and b differ in meaningfulness, of course, but they also differ in familiarity. We simply have more experience processing bs as compared with ꟼs. In a very simple study, Lupyan and Spivey (2008) used a visual search task in which participants were asked to search for a ⊓ among ⊔s (or vice-versa). The stimuli were meaningless and perceptually novel. Some participants were explicitly told at the start of the experiment that the shapes should be thought of as rotated 2 and 5s. This simple instruction dramatically improved overall RTs and led to shallower search slopes, indicating more efficient visual processing. The effect of construing a stimulus as meaningful (and in this case, associating it with a named category) produced a sustained effect in the sense that once induced, the facilitation persists, an effect reminiscent of the well-known hidden Dalmatian in a piecemeal image, which once known to be present in the image, cannot be "un-seen" (Gregory, 1970; see also Porter, 1954). Arguably, such effects are *also* on-line effects (see also Bentin and Golland, 2002). The degree to which such conceptual effects on visual processing are truly linguistic requires further investigation and neurostimulation techniques such as TMS and tDCS will potentially prove useful (Lupyan et al., 2010a).

These results potentially inform the findings of cross-linguistic differences in early ERPs in response to changing colors. Thierry et al. (2009) found that Greek speakers who, like Russian speakers, have separate words for light and dark blues, showed a greater visual mismatch negativity – an early component showing condition-differences starting at ∼160 ms that has been used to index automatic, and arguably preattentive change detection – when presented with color changes that spanned the lexical boundary. The authors found some differences in the P1 component as well. On the one hand, such differences in early visual processing may be viewed as consequences of long-term perceptual warping produced by language (or perhaps other cultural factors). This account however, would be at a loss to explain why in other studies verbal interference can eliminate cross-linguistic differences on behavioral measures of categorical color perception. An alternative account is that viewing colors automatically activates their names that warp perceptual representations on-line. The observed effects on early perception are thus evidence not of a permanent change in bottom-up processing, but rather of a sustained top-down modulation possibly induced by activation of the color names during the task[5].

THE NEURAL PLAUSIBILITY OF LANGUAGE-MODULATED PERCEPTION

Understanding the word "chair" is clearly a more complex process than detecting the presence of a shape in a visual display or determining which of two color swatches matches a third. How can a complex "high-level" process influence low-level and much more rapid processes such as simple detection? This would indeed be puzzling if the brain were a feedforward system. It is not. Neural processing is intrinsically interactive (Mesulam, 1998; Freeman, 2001). As eloquently argued in a prescient paper by Churchland et al. (1994), the brain is only grossly hierarchical: sensory input signals are only a part of what drives "sensory" neurons, processing stages are not like assembly line productions, and later processing can influence earlier processing (p. 59). This view has in recent years received overwhelming support (e.g., Mumford, 1992; Rao and Ballard, 1999; Lamme and Roelfsema, 2000; Foxe and Simpson, 2002; Reynolds and Chelazzi, 2004; Gilbert and Sigman, 2007; Kveraga et al., 2007; Mesulam, 2008; Koivisto et al., 2011).

To give two examples from vision of gross violations of hierarchical processing: (1) the "late" prefrontal areas of cortex can at times respond to the presence of a visual stimulus *before* early visual cortex (V2; Lamme and Roelfsema, 2000 for review). (2) The well-known classical receive fields of V1 neurons showing orientation tuning appear to be dynamically reshaped by horizontal and top-down processes. Within 100 ms. after stimulus onset, V1 neurons are re-tuned from reflecting simple orientation features, to representing figure/ground relationships over a much larger visual angle (Olshausen et al., 1993; Lamme et al., 1999).

Effects of verbal labels on vision can be seen as embodying a similar, but more complex type of perceptual modulation as the reshaping of V1 receptive fields. Although the neural loci of these effects are at present unknown, one possibility is that processing an object name initiates a volley of feedback activity to object-selective regions of cortex such as IT (Logothetis and Sheinberg, 1996), producing a predictive signal or "head start" to the visual system (Kveraga et al., 2007; Esterman and Yantis, 2008; Puri and Wojciulik, 2008). On several theories of attention (e.g., biased competition theory of Desimone and Duncan, 1995), these predictive signals would enable neurons that respond to the named object to gain a competitive advantage (see also Vecera and Farah, 1994; Kramer et al., 1997; Deco and Lee, 2002; Kravitz and Behrmann, 2008). Given feedback from object-selective cortical regions, winning objects can bias earlier spatial regions of visual cortex.

LABELS AND STIMULUS TYPICALITY

The two-dimensional projection of the perceptual representations shown in **Figure 2** hides an interesting interaction between labels and stimulus typicality. Not surprisingly, the network shows basic typicality effects. The correct category label is more quickly and/or strongly activated when the network is presented with a more typical item (i.e., an item having more typical values on dimensions learned by the network to be important). The somewhat counter-intuitive consequence is that it is these already typical items that are most affected by labels: the items tend to become even more typical as the network fills in undefined or unknown features with category-typical values. The atypical exemplars (i.e., instances on the periphery of the category), although having the most *potential* to be affected by the label, interact with the label more weakly than the more central exemplars. One can visualize this effect using a magnet metaphor: an object positioned far from a magnet can

[5]The authors did not test whether linguistic manipulations such as verbal interference reduce or eliminate the cross-linguistic difference in the visual mismatch negativity, although in a commentary they admit that this would be a natural followup (Athanasopoulos et al., 2009).

be moved a greater distance than an object positioned close to the magnet, but because the magnetic field drops off rapidly with increasing distance, the object farther away is being pulled only weakly and may not move at all. Such a mechanism has similarities to the perceptual magnet effect in perception of phonemes (Kuhl, 1994) and the attractor field model in visual perception (e.g., Tanaka and Corneille, 2007).

Effects of typicality turn out to be quite pervasive: In visual tasks, up-regulating the effect of labels through overt presentation of the label benefits typical category members more than atypical ones. Effects of labels on perceptual processing appear to be stronger for more typical exemplars. For instance, the effect of hearing a label is strong for a numeral in a typical font (5), compared to when it was rendered in a less typical font (5; Lupyan, 2007; Lupyan and Spivey, 2010b). In the recognition memory task described above (Lupyan, 2008b) it labeling the typical exemplars led to poorer memory whereas labeling atypical exemplars did not. As a further demonstration that processing an item in the context of its name activates a more typical representation, consider the following two results:

(1) In Experiment 6 of Lupyan (2008b), participants were asked to rate pictures of chairs and lamps on typicality (from very typical to very atypical). The pictures were presented, one at a time, followed by a prompt with the rating scale. The text of the prompt either mentioned the name of the category by name ("chair"/"lamp") or did not (a within-subject manipulation). Participants were instructed to always rate the object's typicality with respect to its category. That is, the task was the same regardless of how the prompt was worded. Yet, participants were more likely to rate the same pictures as more typical when asked, "How typical was that chair" than "How typical was that object," rating the already typical objects *more* typical when referred to by their name (**Figure 3**).

(2) Categories like chair, although comprising concrete objects, are rather fuzzy and do not have formal definitions. In contrast, categories like triangle, *can* be formally defined (Armstrong et al., 1983). All triangles are three-sided polygons and all three-sided polygons are triangles. When queried, all tested participants (18/18) correctly stated this formal definition. When tested on a speeded recognition task, participants showed a typicality/canonicality effect, being faster to recognize isosceles than scalene triangles. This effect, however, was obtained only on trials when participants were cued with the word "triangle." When, on randomly intermixed trials, participants were cued with the phrase "three sides," they were equally fast to recognize isosceles and scalene triangles. According to the label-feedback hypothesis, the category label "triangle" activates a more typical triangle, which in this case appears to correspond to an isosceles/equilateral triangle with a horizontal base. One interesting prediction is that if the label tends to activate a canonical triangle, then referring to a non-canonical triangle explicitly with the word "triangle" may actually alter judgments of its physical properties. To test this prediction, participants were asked to estimate the angle of triangles with a prompt that asked to either estimate the angle of "this triangle" or of "this three-sided figure" (with the instruction varying between-subjects). Participants in the *triangle* condition over-estimated the angle more than participants in the *three-sided* condition (**Figure 4**) – possibly caused by a contrast effect between the activated canonical (non-rotated) triangle and the rotated triangle being judged. This difference persisted for the entire length of the experiment (about 150 trials; Lupyan, 2011; Lupyan et al., in preparation).

THE ROLE OF VERBAL LABELS IN THE LEARNING OF NOVEL CATEGORIES

The learning of categories is in principle separable from the learning of their names. A child, for example, can have a conceptual category of "dog" (such that different dogs are reliably classified as being the same kinds of thing) without having a name for the category. In practice, however, the two processes are intimately linked. Not only does conceptual development

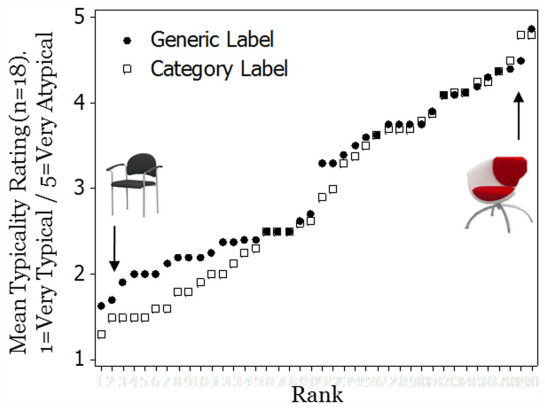


FIGURE 3 | A comparison of typicality ratings of chairs and lamps when the prompt includes the category name ("chair" or "lamp") and when it includes a generic referent ("object"; Lupyan, 2008b, Experiment 6). The x-axis is the average typicality rank of each picture from very typical to very atypical.

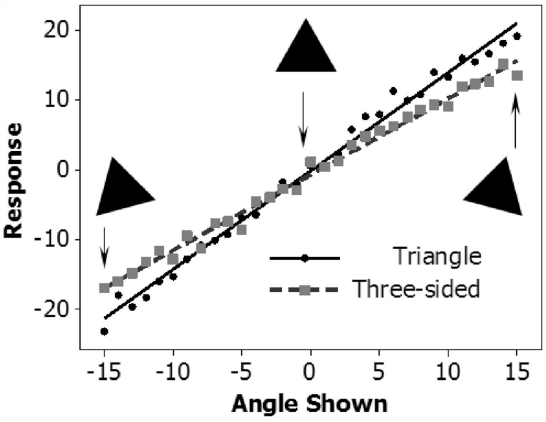


FIGURE 4 | A comparison of judgments of the base angle relative to the horizontal of triangles called "triangles" and the same figures called "three-sided shapes" (Lupyan, 2011).

shape verbal development (e.g., Snedeker and Gleitman, 2004), but verbal learning impacts conceptual development (Waxman and Markow, 1995; Gumperz and Levinson, 1996; Levinson, 1997; Spelke and Tsivkin, 2001; Gentner and Goldin-Meadow, 2003; Yoshida and Smith, 2005; Lupyan et al., 2007). The idea that language shapes concepts has two implications. The first is that it is of course *through* language that we learn much of what we know. This is often seen as trivial, as when Devitt and Sterelny wrote, apparently without irony, that "the only respect in which language clearly and obviously does influence thought turns out to be rather banal: language provides us with most of our concepts" (Devitt and Strelny, 1987, p. 178)[6]. The second implication is that the very use of words may facilitate, or in some cases enable, the ability to impose categories on the external world. Do category names actually facilitate the learning of novel categories?

In a study designed to answer this question, Lupyan et al. (2007) compared the ability of participants to learn categories that were labeled to the learning of the same categories without names. The basic task required participants to learn to classify 16 "aliens" into those that ought to be approached and those to be avoided, responding with the appropriate direction of motion (approach/escape). The category distinction involved subtle differences in the configuration of the "head" and "body" of the creatures. On each training trial, one of the 16 aliens appeared in the center of the screen and had to be categorized by moving a character in a spacesuit (the "explorer") toward or away from the alien, with auditory feedback marking the response as correct or not. In the *label* conditions, a printed or auditory label (the nonsense terms, "leebish" and "grecious") appeared next to the alien; in the *no-label* condition, the alien remained on the screen by itself. All the participants received the same number of categorization trials and saw the aliens for exactly the same duration; the only difference between the groups was the presence of the category labels that followed each response. The labels, being perfectly predictive of the behavioral responses, constituted entirely redundant information.

The results showed that participants in the label conditions learned to classify the aliens much faster than those in the no-label conditions. When the labels were replaced with equally redundant and easily learned non-linguistic and non-referential cues (corresponding to where the alien lived), the cues failed to facilitate categorization. After completing the category-training phase during which participants in both groups eventually reached ceiling performance, their knowledge of the categories was tested in a speeded categorization task using a combination of previously categorized and novel aliens, presented without any feedback or labels. Results showed that those who learned the categories in the presence of labels retained their category knowledge throughout the testing phase. Those who learned the categories without labels showed a decrease in accuracy over time. Thus, learning named categories appears to be easier than learning unnamed categories. More than just learning to map words onto pre-existing concepts (cf. Li and Gleitman, 2002; Snedeker and Gleitman, 2004), words appear to facilitate the categorization process itself.

IMPLICATIONS OF THE LABEL-FEEDBACK HYPOTHESIS FOR THE "LANGUAGE AND THOUGHT" RESEARCH PROGRAM

Most work investigating the relationship between language, cognition, and perception has assumed that verbal and non-verbal representations are fundamentally distinct and the goal of the "language and thought" research program is to understand whether and how linguistic representations affect non-linguistic representations (Wolff and Holmes, 2011). On such a view, information communicated or encoded via language comprises what is essentially a separate "verbal" modality or channel (Paivio, 1986). Linguistic effects are ascribed either to language influencing "deep" non-verbal processes which ought to not be affected by verbal interference or acquired language deficits, or else hinge on high-level processes that combine verbal and non-verbal input in some way (e.g., Roberson and Davidoff, 2000; Pilling et al., 2003; Dessalegn and Landau, 2008; Mitterer et al., 2009). Neither proposed mechanism can explain how language can have pervasive effects on perceptual processing that are nevertheless permeable to linguistic manipulations such as verbal interference – the paradox outlined above.

The label-feedback hypothesis provides a way of resolving the paradox. Effects of language can indeed run "deep" in the sense of affecting low-level processes (e.g., Thierry et al., 2009) – the very processes claimed by Gleitman (2010) to be impervious to language. Such effects of language on, e.g., color perception need not arise from language somehow permanently warping perceptual space. Thinking of these effects as occurring on-line explains why they can be modulated by verbal factors such as overt language use and verbal interference. Framing effects of language as occurring on-line does not render them superficial, strategic, or necessarily under voluntary control (Lupyan and Spivey, 2010a,b; Lupyan et al., 2010b). On this formulation, the distinction between verbal and non-verbal representations becomes moot, just as taking seriously the pervasiveness of top-down effects in perception renders moot the distinction between "earlier" and "later" cortical areas (Gilbert and Sigman, 2007).

To return to the case of linguistic effects on color perception: On the present view, a visual representation of a color, e.g., blue, becomes rapidly modulated by the activation of the word "blue," a process that can be exaggerated by exogenous presentation of the label and attenuated by manipulations such as verbal interference. Thus, although the *bottom-up* processing of color is likely to be independent of language and identical in speakers of different languages, the *top-down* effects in which language takes part are dependent on the word-color associations to which the speakers have been exposed, and will thus be correspondingly different between speakers who possess a generic term "blue" and those who do not. Such modulations occur as the label becomes active (over the course of a few 100 ms). There is nothing mysterious about this process: it is simply the consequence of the idea that visual representations involved in making even the simplest visual decisions are augmented by feedback higher-level, and typically more anterior brain regions. Feedback from language-based activations such as the activation of the word "green" on seeing green color patches can be seen as one form of such top-down influence.

[6]It is probably too obvious to mention, but this function of language is far from banal. Consider that in the absence of language, much of what humans need to learn to survive would have to be learned through slow and dangerous trial and error (Harnad, 2005). It is not an exaggeration to claim that without the ability to learn through language human culture would not exist (Deacon, 1997).

Although color processing has been a popular testing ground for exploring effects of language[7], the label-feedback hypothesis has a broader relevance. At stake is the question of whether and to what degree perception of familiar objects is continuously augmented by the labels that become co-active with perceptual representations of these objects. This means that once a label is learned, it can potentially modulate subsequent processing (visual and otherwise) of objects to which the label refers. Indeed, the benefits of names in learning novel categories (Lupyan et al., 2007), may derive, at least in part, from the labels' effect on perceptual processing of the exemplars (see also Lupyan and Thompson-Schill, 2012). Lexicalization patterns differ substantially between languages (e.g., Bowerman and Choi, 2001; Lucy and Gaskins, 2001; Majid et al., 2007; Evans and Levinson, 2009). Accordingly, speakers of different languages end up with different patterns of associations between labels and external objects, resulting in different top-down effects of language on ongoing "non-verbal" processing in speakers of different languages.

The label-feedback hypothesis as presented here does not claim to be relevant to all effects labeled as "Whorfian" in the literature. The most direct application is to the processes of categorization and object perception. The hypothesis does not predict that any differences in the grammar of language translate to meaningful differences in "thought." A pervasive additional source of confusion in the language and thought literature that I have not discussed here relates to predicting the consequences that a particular linguistic difference should have on a particular putatively non-linguistic task. Consider, for example, the observation that English verbs highlight the manner of motion (e.g., walk, run, hop) leaving the path as an option, while Spanish verbs highlight the path of motion (e.g., entrar, pasar) leaving the manner as an option (Talmy, 1988). Does the priority of manner information in English mean that English speakers should have better memory for manner than Spanish speakers? Perhaps, but one might just as easily predict the opposite pattern: Spanish speakers ought to have better memory for manner information because, when it is mentioned, it is more unexpected and thus more salient (cf. Gennari et al., 2002; Papafragou et al., 2008). Progress in this area appears to require a firmer marriage between memory researchers and psycholinguists.

More generally, rather than attempting to decide whether a given representation comprises a verbal or visual "code" (e.g., Dessalegn and Landau, 2008), on the current proposal, it may be more productive to measure the degree to which performance on *specific tasks* is being modulated by language, modulated differently by different languages, or is truly independent of any experimental manipulations that can be termed linguistic. On this account, the central question is not "do speakers of different languages have different color concepts" but rather "how does language affect the perceptual representations of color brought to bear on a given task." Much of the literature in the language and thought arena holds an implicit (and sometimes explicit) assumption that there exists such things as *the* concept of a dog, or *the* concept of green-ness. On this assumption, accepting that the concept of green-ness is influenced by language creates the expectation that one should observe those linguistic effects on any task that taps into that singular color concept. Failure to observe these effects is then used by as an argument against linguistic relativity or language-mediated vision. On an alternative view, however, conceptual representations are dynamic assemblies that are a function of prior knowledge as well as current task demands (Casasanto and Lupyan, 2011; Lupyan et al., under review; see also Prinz, 2004). There is therefore no single concept of green-ness. Rather, the influence of language on ongoing cognitive and perceptual processing may be present in some tasks and non-existent in others. For example, given the categorical nature of linguistic reference, one prediction is that effects of language ought to become stronger in tasks that require or promote categorization and weaker in tasks that discourage it (e.g., realistic drawing, remembering exact spatial locations, judging a continuously varying motion trajectory). By understanding how language may augment specific cognitive and perceptual processes, we can make predictions about the kinds of tasks should or should not be influenced by language broadly construed and by differences between languages.

CONCLUSION

I have argued that a pervasive source of theoretical confusion regarding effects of language on cognition and perception stems from a failure to appreciate the degree to which virtually all cognitive and perceptual acts reflect interactive-processing, combining bottom-up and top-down sources of information. An effect of language on how we perceive the rainbow does not require it to alter the responses of photoreceptors. A deep and persistent effect of language on object concepts does not require it to alter conceptual "cores" (indeed, the very existence of such conceptual cores is debatable, Barsalou, 1987; Prinz, 2004; Casasanto and Lupyan, 2011).

Our perception of rainbows, dogs, and everything in between is a product of both their physical properties and top-down processes. The idea that words affect ongoing cognitive and perceptual processes via top-down feedback provides a useful way for thinking about the interaction of language with other processes. In its present form, the label-feedback hypothesis is merely a sketch, but as evidenced by some of the studies reviewed in this paper, this framework provides a powerful intuition pump for generating testable predictions. The label-feedback hypothesis is broadly consistent with what we know about neural mechanisms of perception and categorization, although its neural underpinnings remain almost completely unexplored. The next step is to understand these mechanisms.

ACKNOWLEDGMENTS

Portions of this manuscript appeared in Lupyan (2007). *The Label-Feedback Hypothesis: Linguistic Influences on Visual Processing.* PhD. Thesis. Carnegie Mellon University, during which time the author was supported by an NSF Graduate fellowship.

[7]Witzel and Gegenfurtner (2011) present a cogent argument that most recent investigations of categorical color perception have made incorrect assumptions regarding psychophysical distances in the CIE color space, such that color pairs claimed to be equally spaced in psychophysical space may not be, rendering many of the claims made by these studies difficult to interpret.

REFERENCES

Armstrong, S. L., Gleitman, L. R., and Gleitman, H. (1983). What some concepts might not be. *Cognition* 13, 263–308.

Athanasopoulos, P., Wiggett, A., Dering, B., Kuipers, J.-R., and Thierry, G. (2009). The Whorfian mind: electrophysiological evidence that language shapes perception. *Commun. Integr. Biol.* 2, 332–334.

Barsalou, L. W. (1987). "The instability of graded structure: implications for the nature of concepts," in *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, ed. U. Neisser (Cambridge: Cambridge University Press), 101–140.

Bentin, S., and Golland, Y. (2002). Meaningful processing of meaningless stimuli: the influence of perceptual experience on early visual processing of faces. *Cognition* 86, B1–B14.

Black, M. (1959). Linguistic relativity: the views of Benjamin Lee Whorf. *Philos. Rev.* 68, 228–238.

Borges, J. L. (1942/1999). *Collected Fictions.* Penguin (New York, NY: Non-Classics).

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cogn. Psychol.* 43, 1–22.

Boroditsky, L. (2010). "How the languages we speak shape the ways we think: the FAQs," in *The Cambridge Handbook of Psycholinguistics*, eds M. J. Spivey, M. Joanisse, and K. McRae (Cambridge: Cambridge University Press).

Boroditsky, L. (2003). "Linguistic relativity," in *Encyclopedia of Cognitive Science*, ed. L. Nadel (London: Macmillan), 917–922.

Bowerman, M., and Choi, S. (2001). "Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories," in *Language Acquisition and Conceptual Development*, eds M. Bowerman and S. C. Levinson (Cambridge: Cambridge University Press), 475–511.

Carruthers, P. (2002). The cognitive functions of language. *Behav. Brain Sci.* 25, 657–674.

Casasanto, D. (2008). Who's afraid of the big bad whorf? Crosslinguistic differences in temporal language and thought. *Lang. Learn.* 58, 63–79.

Casasanto, D., and Lupyan, G. (2011). Ad hoc cognition. *Presented at the Annual Conference of the Cognitive Science Society*, Boston, MA.

Cassirer, E. (1962). *An Essay on Man: An Introduction to a Philosophy of Human Culture.* New Haven, CT: Yale University Press.

Churchland, P. S., Ramachandran, V., and Sejnowski, T. J. (1994). "A critique of pure vision," in *Large-Scale Neuronal Theories of the Brain*, eds C. Koch and J. L. Davis (Cambridge, MA: The MIT Press), 23–60.

Clark, A. (1998). "Magic words: how language augments human computation," in *Language and Thought: Interdisciplinary Themes*, eds P. Carruthers and J. Boucher (New York, NY: Cambridge University Press), 162–183.

Cohen, R., Kelter, S., and Woll, G. (1980). Analytical competence and language impairment in aphasia. *Brain Lang.* 10, 331–347.

Cohen, R., Woll, G., Walter, W., and Ehrenstein, H. (1981). Recognition deficits resulting from focussed attention in aphasia. *Psychol. Res.* 43, 391–405.

Daoutis, C. A., Franklin, A., Riddett, A., Clifford, A., and Davies, I. R. L. (2006). Categorical effects in children's colour search: a cross-linguistic comparison. *Br. J. Dev. Psychol.* 24, 373–400.

Davidoff, J., Davies, I. R. L., and Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature* 398, 203–204.

Davidoff, J., and Roberson, D. (2004). Preserved thematic and impaired taxonomic categorisation: a case study. *Lang. Cogn. Process.* 19, 137–174.

Davies, I. R. L., and Corbett, G. (1998). A cross-cultural study of color-grouping: tests of the perceptual-physiology account of color universals. *Ethos* 26, 338–360.

Deacon, T. (1997). *The Symbolic Species: The Co-Evolution of Language and the Brain.* London: The Penguin Press.

Deco, G., and Lee, T. S. (2002). A unified model of spatial and object attention based on inter-cortical biased competition. *Neurocomputing* 44, 775–781.

Dennett, D. C. (1994). "The role of language in intelligence," in *What is Intelligence? The Darwin College Lectures*, ed. J. Khalfa (Cambridge: Cambridge University Press).

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual-attention. *Annu. Rev. Neurosci.* 18, 193–222.

Dessalegn, B., and Landau, B. (2008). More than meets the eye: the role of language in binding and maintaining feature conjunctions. *Psychol. Sci.* 19, 189–195.

Devitt, M., and Strelny, K. (1987). *Language and Reality: An Introduction to the Philosophy of Language.* Cambridge, MA: MIT Press.

Drivonikou, G. V., Kay, P., Regier, T., Ivry, R. B., Gilbert, A. L., Franklin, A., and Davies, I. R. L. (2007). Further evidence that Whorfian effects are stronger in the right visual field than the left. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1097–1102.

Esterman, M., and Yantis, S. (2008). Category expectation modulates object-selective cortical activity. *J. Vis.* 8, 555a.

Evans, N., and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32, 429.

Fausey, C. M., and Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychon. Bull. Rev.* 18, 150–157.

Folstein, J. R., Gauthier, I., and Palmeri, T. J. (2010). Mere exposure alters category learning of novel objects. *Front. Psychol.* 1:40. doi:10.3389/fpsyg.2010.00040

Foxe, J. J., and Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans – a framework for defining "early" visual processing. *Exp. Brain Res.* 142, 139–150.

Freeman, W. J. (2001). *How Brains Make Up Their Minds*, 1st Edn. New York, NY: Columbia University Press.

Gennari, S. P., Sloman, S. A., Malt, B. C., and Fitch, W. T. (2002). Motion events in language and cognition. *Cognition* 83, 49–79.

Gentner, D., and Goldin-Meadow, S. (2003). *Language in Mind: Advances in the Study of Language and Thought.* Cambridge, MA: MIT Press.

Gentner, D., and Loewenstein, J. (2002). "Relational language and relational thought," in *Language, Literacy, and Cognitive Development*, eds J. Byrnes and E. Amsel (Mahwah, NJ: LEA), 87–120.

Gilbert, A. L., Regier, T., Kay, P., and Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proc. Natl. Acad. Sci. U.S.A.* 103, 489–494.

Gilbert, C. D., and Sigman, M. (2007). Brain states: top-down influences in sensory processing. *Neuron* 54, 677–696.

Gleitman, H., Fridlund, A. J., and Reisberg, D. (2004). *Psychology*, 6th Edn. New York: Norton and Company.

Gleitman, L. (2010). *Economist Debates: Language: This House Believes that the Language We Speak Shapes How We Think.* Available at: http://www.economist.com/debate/days/view/632 [retrieved September 1, 2011].

Gleitman, L., and Papafragou, A. (2005). "Language and thought," in *Cambridge Handbook of Thinking and Reasoning*, eds K. Holyoak and B. Morrison (Cambridge: Cambridge University Press), 633–661.

Goldstein, K. (1924/1948). *Language and Language Disturbances.* New York: Grune and Stratton.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *J. Exp. Psychol. Gen.* 123, 178–200.

Goldstone, R. L. (1998). Perceptual learning. *Ann. Rev. Psychol.* 49, 585–612.

Goldstone, R. L., and Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition* 65, 231–262.

Goldstone, R. L., and Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 69–78.

Goldstone, R. L., Lippa, Y., and Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition* 78, 27–43.

Goldstone, R. L., Steyvers, M., and Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Mem. Cognit.* 31, 169–180.

Gopnik, A. (2001). "Theories, language, and culture: whorf without wincing," in *Language Acquisition and Conceptual Development*, eds M. Bowerman and S. C. Levinson (Cambridge: Cambridge University Press), 45–69.

Gregory, R. L. (1970). *The Intelligent Eye.* New York: McGraw-Hill.

Gumperz, J. J., and Levinson, S. C. (1996). *Rethinking Linguistic Relativity.* Cambridge: Cambridge University Press.

Harnad, S. (2005). "Cognition is categorization," in *Handbook of Categorization in Cognitive Science*, eds H. Cohen and C. Lefebvre (San Diego, CA: Elsevier Science), 20–45.

James, W. (1890). *Principles of Psychology*, Vol. 1. New York: Holt.

January, D., and Kako, E. (2007). Re-evaluating evidence for linguistic relativity: reply to Boroditsky (2001). *Cognition* 104, 417–426.

Kemmerer, D. (2010). "How words capture visual experience: the perspective from cognitive neuroscience," in *Words and the Mind: How Words Capture Human Experience*, 1st Edn, eds B. Malt and P. Wolff (New York, NY: Oxford University Press), 289–329.

Kemmerer, D., Rudrauf, D., Manzel, K., and Tranel, D. (2010). Behavioral patterns and lesion sites associated with impaired processing of lexical and conceptual knowledge of actions. *Cortex*. doi:10.1016/j.cortex.2010.11.001. [Epub ahead of print].

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., and Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *J. Neurosci.* 31, 2488–2492.

Kotovsky, L., and Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Dev.* 67, 2797–2822.

Kramer, A. F., Weber, T. A., and Watson, S. E. (1997). Object-based attentional selection – grouped arrays or spatially invariant representations?: comment on Vecera and Farah (1994). *J. Exp. Psychol. Gen.* 126, 3–13.

Kravitz, D. J., and Behrmann, M. (2008). The space of an object: object attention alters the spatial gradient in the surround. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 298–309.

Kuhl, P. (1994). Learning and representation in speech and language. *Curr. Opin. Neurobiol.* 4, 812–822.

Kveraga, K., Ghuman, A. S., and Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain Cogn.* 65, 145–168.

Lamme, V. A. F., Rodriguez-Rodriguez, V., and Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cereb. Cortex* 9, 406–413.

Lamme, V. A. F., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579.

Lee, P. (1996). *The Whorf Theory Complex: A Critical Reconstruction.* Philadelphia, PA: John Benjamins Pub Co.

Levinson, S. C. (1997). "From outer to inner space: linguistic categories and non-linguistic thinking," in *Language and Conceptualization*, eds J. Nuyts and E. Pederson (Cambridge: Cambridge University Press), 13–45.

Li, P., Dunham, Y., and Carey, S. (2009). Of substance: the nature of language effects on entity construal. *Cogn. Psychol.* 58, 487–524.

Li, P., and Gleitman, L. (2002). Turning the tables: language and spatial reasoning. *Cognition* 83, 265–294.

Liu, Q., Li, H., Campos, J. L., Wang, Q., Zhang, Y., Qiu, J., Zhang, Q., and Sun, H. J. (2009). The N2pc component in ERP and the lateralization effect of language on color perception. *Neurosci. Lett.* 454, 58–61.

Loewenstein, J., and Gentner, D. (2005). Relational language and the development of relational mapping. *Cogn. Psychol.* 50, 315–353.

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621.

Lucy, J. A., and Gaskins, S. (2001). "Grammatical categories and the development of classification preferences: a comparative approach," in *Language Acquisition and Conceptual Development*, eds M. Bowerman and S. C. Levinson (Cambridge: Cambridge University Press), 257–283.

Lupyan, G. (2007). "Reuniting categories, language, and perception," in *Twenty-Nineth Annual Meeting of the Cognitive Science Society*, eds D. S. McNamara and J. G. Trafton (Austin, TX: Cognitive Science Society), 1247–1252.

Lupyan, G. (2008a). The conceptual grouping effect: categories matter (and named categories matter more). *Cognition* 108, 566–577.

Lupyan, G. (2008b). From chair to "chair:" a representational shift account of object labeling effects on memory. *J. Exp. Psychol. Gen.* 137, 348–369.

Lupyan, G. (2009). Extracommunicative functions of language: verbal interference causes selective categorization impairments. *Psychon. Bull. Rev.* 16, 711–718.

Lupyan, G. (2011). Representations of basic geometric shapes are created ad-hoc. concepts, actions, and objects workshop. *Presented at the Concepts, Actions, and Objects Workshop*, Rovereto.

Lupyan, G. (in press). "What do words do? Towards a theory of language-augmented thought," in *The Psychology of Learning and Motivation*, Vol. 57, ed. B. H. Ross.

Lupyan, G., Mirman, D., Hamilton, R. H., and Thompson-Schill, S. L. (2010a). Linking language, cognitive control, and categorization: evidence from aphasia and transcranial direct current stimulation. *Presented at the Second Conference on the Neurobiology of Language*, San Diego, CA.

Lupyan, G., Thompson-Schill, S. L., and Swingley, D. (2010b). Conceptual penetration of visual processing. *Psychol. Sci.* 21, 682–691.

Lupyan, G., Rakison, D. H., and McClelland, J. L. (2007). Language is not just for talking: labels facilitate learning of novel categories. *Psychol. Sci.* 18, 1077–1082.

Lupyan, G., and Spivey, M. J. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli. *Curr. Biol.* 18, R410–R412.

Lupyan, G., and Spivey, M. J. (2010a). Making the invisible visible: auditory cues facilitate visual object detection. *PLoS ONE* 5, e11452. doi:10.1371/journal.pone.0011452

Lupyan, G., and Spivey, M. J. (2010b). Redundant spoken labels facilitate perception of multiple items. *Atten. Percept. Psychophys.* 72, 2236–2253. doi:10.3758/APP.72.8.2236

Lupyan, G., and Swingley, D. (in press). Self-directed speech affects visual processing. *Q. J. Exp. Psychol.*

Lupyan, G., and Thompson-Schill, S. L. (2012). The evocative power of words: activation of concepts by verbal and nonverbal means. *J. Exp. Psychol. Gen.* 141, 170–186.

Macpherson, F. (2012). Cognitive penetration of colour experience: rethinking the issue in light of an indirect mechanism. *Philos. Phenomenol. Res.* 84, 24–62.

Majid, A., Gullberg, M., van Staden, M., and Bowerman, M. (2007). How similar are semantic categories in closely related languages? A comparison of cutting and breaking in four Germanic languages. *Cogn. Linguist.* 18, 179–194.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception 1. An account of basic findings. *Psychol. Rev.* 88, 375–407.

Mesulam, M. M. (1998). From sensation to cognition. *Brain* 121, 1013–1052.

Mesulam, M. M. (2008). Representation, inference, and transcendent encoding in neurocognitive networks of the human brain. *Ann. Neurol.* 64, 367–378.

Meteyard, L., Bahrami, B., and Vigliocco, G. (2007). Motion detection and motion verbs – language affects low-level visual perception. *Psychol. Sci.* 18, 1007–1013.

Mitterer, H., Horschig, J. M., Musseler, J., and Majid, A. (2009). The influence of memory on perception: it's not what things look like, it's what you call them. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1557–1562.

Mo, L., Xu, G., Kay, P., and Tan, L.-H. (2011). Electrophysiological evidence for the left-lateralized effect of language on preattentive categorical perception of color. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14026–14030.

Mumford, D. (1992). On the computational architecture of the neocortex II. The role of cortico-cortical loops. *Biol. Cybern.* 66, 251.

Noppeney, U., and Wallesch, C. W. (2000). Language and cognition – Kurt Goldstein's theory of semantics. *Brain Cogn.* 44, 367–386.

Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* 13, 4700–4719.

Ozgen, E., and Davies, I. R. L. (2002). Acquisition of categorical color perception: a perceptual learning approach to the linguistic relativity hypothesis. *J. Exp. Psychol. Gen.* 131, 477–493.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. New York: Oxford University Press.

Papafragou, A., Hulbert, J., and Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition* 108, 155–184.

Pilling, M., Wiggett, A., Ozgen, E., and Davies, I. R. L. (2003). Is color "categorical perception" really perceptual? *Mem. Cognit.* 31, 538–551.

Porter, P. B. (1954). Another picture puzzle. *Am. J. Psychol.* 67, 550–551.

Prinz, J. J. (2004). *Furnishing the Mind: Concepts and their Perceptual Basis.* The MIT Press.

Puri, A. M., and Wojciulik, E. (2008). Expectation both helps and hinders object perception. *Vision Res.* 48, 589–597.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav. Brain Sci.* 22, 341–365.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nat. Neurosci.* 2, 79–87.

Ratterman, M. J., and Gentner, D. (1998). "The effect of language on similarity: the use of relational symbols improves young children's performance on a mapping task," in *Advances in Analogy Research: Integration of Theory and Data From the Cognitive, Computational and Neural Sciences*, eds K. Holyoak, D. Gentner and B. Kokinov (Sophia: New Bulgarian University).

Reynolds, J. H., and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annu. Rev. Neurosci.* 27, 611–47.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., and Brecher, A. (1996). The Philadelphia naming test: scoring and rationale. *Am. J. Speech Lang. Pathol.* 24, 121–133.

Roberson, D., and Davidoff, J. (2000). The categorical perception of colors and facial expressions: the effect of verbal interference. *Mem. Cognit.* 28, 977–986.

Roberson, D., Davidoff, J., Davies, I. R. L., and Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. *Cogn. Psychol.* 50, 378–411.

Roberson, D., Pak, H., and Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: evidence from Korean. *Cognition* 107, 752–762.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vols 1 and 2. Cambridge, MA: MIT Press.

Slobin, D. (1996). "From 'thought and language' to 'thinking for speaking,'" in *Rethinking Linguistic Relativity*, eds J. J. Gumperz and S. C. Levinson (Cambridge: Cambridge University Press), 70–96.

Sloutsky, V. M. (2010). From perceptual categories to concepts: what develops? *Cogn. Sci.* 34, 1244–1286.

Snedeker, J., and Gleitman, L. (2004). "Why is it hard to label our concepts?" in *Weaving a Lexicon*, eds D. G. Hall and S. R. Waxman (Cambridge, MA: The MIT Press), 257–294.

Spelke, E. S. (2003). "What makes us smart? Core knowledge and natural language," in *Language in Mind: Advances in the Study of Language and Thought*, eds D. Gentner and S. Goldin-Meadow (Cambridge, MA: MIT Press), 277–311.

Spelke, E. S., and Tsivkin, S. (2001). "Initial knowledge and conceptual change: space and number," in *Language Acquisition and Conceptual Development* (Cambridge Cambridge University Press), 475–511.

Talmy, L. (1988). Force dynamics in language and cognition. *Cogn. Sci.* 12, 49–100.

Tan, L. H., Chan, A. H. D., Kay, P., Khong, P.-L., Yip, L. K. C., and Luke, K.-K. (2008). Language affects patterns of brain activation associated with perceptual decision. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4004–4009.

Tanaka, J. W., and Corneille, O. (2007). Typicality effects in face and object perception: further evidence for the attractor field model. *Percept. Psychophys.* 69, 619–627.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., and Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proc. Natl. Acad. Sci. U.S.A.* 106, 4567–4570.

Vecera, S. P., and Farah, M. J. (1994). Does visual-attention select objects or locations. *J. Exp. Psychol. Gen.* 123, 146–160.

Vygotsky, L. (1962). *Thought and Language*. Cambridge, MA: MIT Press.

Ward, E. J., and Lupyan, G. (2011). Linguistic penetration of suppressed visual representations. *Presented at the Vision Sciences Society*, Naples, FL.

Waxman, S. R., and Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cogn. Psychol.* 29, 257–302.

Whorf, B. L. (1956). *Language, Thought, and Reality*. Cambridge, MA: MIT Press.

Wiggett, A., and Davies, I. R. L. (2008). The effect of Stroop interference on the categorical perception of color. *Mem. Cognit.* 36, 231.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7780–7785.

Witzel, C., and Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? *J. Vis.* 11, 16.

Wolff, P., and Holmes, K. (2011). Linguistic relativity. *Wiley Interdiscip. Rev. Cogn. Sci.* 2, 253–265.

Yoshida, H., and Smith, L. B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychol. Sci.* 16, 90–95.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 September 2011; accepted: 11 February 2012; published online: 08 March 2012.

*Citation: Lupyan G (2012) Linguistically modulated perception and cognition: the label-feedback hypothesis. Front. Psychology **3**:54. doi: 10.3389/fpsyg.2012.00054*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 March 2012
doi: 10.3389/fpsyg.2012.00078

How does language change perception: a cautionary note

Nola Klemfuss, William Prinzmetal and Richard B. Ivry*

Department of Psychology, University of California Berkeley, Berkeley, CA, USA

Edited by:
Yury Y. Shtyrov, Medical Research Council, UK

Reviewed by:
Gary Lupyan, University of Wisconsin Madison, USA
Pia Knoeferle, Bielefeld University, Germany

****Correspondence:***
Richard B. Ivry, Department of Psychology, University of California Berkeley, Berkeley, CA 94720-1650, USA.
e-mail: ivry@berkeley.edu

The relationship of language, perception, and action has been the focus of recent studies exploring the representation of conceptual knowledge. A substantial literature has emerged, providing ample demonstrations of the intimate relationship between language and perception. The appropriate characterization of these interactions remains an important challenge. Recent evidence involving visual search tasks has led to the hypothesis that top-down input from linguistic representations may sharpen visual feature detectors, suggesting a direct influence of language on early visual perception. We present two experiments to explore this hypothesis. Experiment 1 demonstrates that the benefits of linguistic priming in visual search may arise from a reduction in the demands on working memory. Experiment 2 presents a situation in which visual search performance is disrupted by the automatic activation of irrelevant linguistic representations, a result consistent with the idea that linguistic and sensory representations interact at a late, response-selection stage of processing. These results raise a cautionary note: While language can influence performance on a visual search, the influence need not arise from a change in perception *per se*.

Keywords: language, perception, embodied cognition, working memory, visual search

INTRODUCTION

Language provides a medium for describing the contents of our conscious experience. We use it to share our perceptual experiences, thoughts, and intentions with other individuals. The idea that language guides our cognition was clearly articulated by Whorf (1956) who proposed that an individual's conceptual knowledge was shaped by his or her language. There is clear evidence demonstrating that language directs thought (Ervin-Tripp, 1967), influences concepts of time and space (e.g., Boroditsky, 2001), and affects memory (e.g., Loftus and Palmer, 1974).

More controversial has been the claim that language has a direct effect on perceptual experience. In a seminal study, Kay and Kempton (1984) found that linguistic labels influence decisions in a color categorization task. In the same spirit, a flurry of studies over the past decade has provided ample demonstrations of how perceptual performance is influenced by language. For example, Meteyard et al. (2007) assessed motion discrimination at threshold for displays of moving dots while participants passively listened to verbs that referred to either motion-related or static actions. Performance on the motion detection task was influenced by the words, with poorer performance observed on the perceptual task when the direction of motion implied by the words was incongruent with the direction of the dot display (see also, Lupyan and Spivey, 2010). Results such as these suggest a close integration of perceptual and conceptual systems (see Goldstone and Barsalou, 1998), an idea captured by the theoretical frameworks of grounded cognition (Barsalou, 2008) and embodied cognition (see Feldman, 2006; Borghi and Pecher, 2011).

There are limitations with tasks based on verbal reports or ones in which the emphasis is on accuracy. In such tasks, language may affect decision and memory processes, as well as perception (see Rosch, 1973). For example, in the Kay and Kempton (1984) study, participants were asked to select the two colored chips that go together best. Even though the stimuli are always visible, a comparison of this sort may engage top-down strategic processes (Pinker, 1997) as well as tax working memory processes as the participant shifts their attentional focus between the stimuli.

To reduce the contribution of memory and decision processes, researchers have turned to simple visual search tasks to explore the influence of language on perception. Consider a visual search study by Lupyan and Spivey (2008). Participants were shown an array of shapes and made speeded responses, indicating if the display was homogeneous or contained an oddball (**Figure 1A**). The shapes were the letters "2" and "5," rotated by 90°. In one condition, the stimuli were described by their linguistic labels. In the other condition, the stimuli were referred to as abstract geometric shapes. RTs were faster for the participants who had been given the linguistic labels or spontaneously noticed that the shapes were rotated letters. Lupyan and Spivey concluded that ". . . visual perception depends not only on what something looks like, but also on what it means" (p. 412).

Visual search has been widely employed as a model task for understanding early perceptual processing (Treisman and Gelade, 1980; Wolfe, 1992). Indeed, we have used visual search to show that the influence of linguistic categories in a detection task is amplified for stimuli presented in the right visual field (Gilbert et al., 2006, 2008). While our results provide compelling evidence that language can influence performance on elementary perceptual tasks, the mechanisms underlying this interaction remain unclear. Lupyan and Spivey (2008; Lupyan, 2008) suggest that the influence of language on perception reflects a dynamic interaction in which linguistic representations sharpen visual feature detectors.

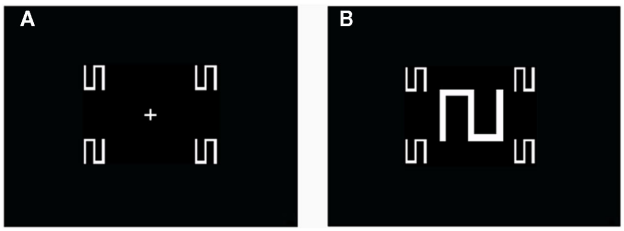


FIGURE 1 | Sample stimulus displays for the No Cue (A) and Cue (B) conditions in Experiment 1. Participants made speeded responses, indicating if the display items were homogenous or contained an oddball. In the Cue conditions, the oddball matched the central cue.

By this view, feedback connections from linguistic or conceptual representations provide a mechanism to bias or amplify activity in perceptual detectors associated with those representations (Lupyan and Spivey, 2010), similar to how attentional cues may alter sensory processing (e.g., Luck et al., 1997; Mazer and Gallant, 2003).

While there is considerable appeal to this dynamic perspective, it is also important to consider alternative hypotheses that may explain how such interactions could arise at higher stages of processing (Wang et al., 1994; Mitterer et al., 2009; see also, Lupyan et al., 2010). Consider the Lupyan and Spivey task from the participants' point of view. The RT data indicate that the displays are searched in a serial fashion (Treisman and Gelade, 1980). When targets are familiar, participants compare each display item to an image stored in long-term memory, terminating the visual search when the target is found. With unfamiliar stimuli, the task is much more challenging (Wang et al., 1994). The participant must form a mental representation of the first shape and maintain this representation while comparing it to each display item. It is reasonable to assume that familiar shapes, ones that can be efficiently coded with a verbal label, would be easier to retain in working memory for subsequent use in making perceptual decisions (Paivio, 1971; Bartlett et al., 1980). In contrast, since unfamiliar stimuli lack a verbal representation in long-term memory, the first item would have to be encoded anew on each trial. We test the memory hypothesis in the following experiment, introducing a condition in which the demands on working memory are reduced.

EXPERIMENT 1

For two groups, the task was similar to that used by Lupyan and Spivey (2008): participants made speeded responses to indicate if a display contained a homogenous set of items or contained one oddball. For two other groups, a cue was present in the center of the display, indicating the target for that trial. Within each display type, one group was given linguistic primes by being told that the displays contained rotated 2's and 5's. The other group was told that the stimuli were abstract forms.

The inclusion of a cue was adopted to minimize the demands on working memory. By pairing the search items with a cue of the target, the task is changed from one requiring an implicit matching process in which each item is compared to a stored representation to one requiring an explicit matching process in which each item is compared to the cue. If language influences perception by priming visual feature detectors, we would expect that participants who were given the linguistic labels would exhibit a similar advantage with both types of displays. In contrast, if the verbal labels reduce the demands on an implicit matching process (e.g., because the verbal labels provide for dual coding in working memory, see Paivio, 1971), then we would expect this advantage to be eliminated or attenuated when the displays contain an explicit cue.

MATERIALS AND METHODS

Participants

Fifty-three participants from the UC Berkeley Research Participation pool were tested. They received class credit for their participation. The research protocol was conducted in accordance with the procedures of the University's Institutional Review Board.

Stimuli

The visual search arrays consisted of 4, 6, or 10 white characters, presented on a black background. The characters were arranged in a circle. The characters were either a "5" or "2," rotated 90°clockwise. The characters fit inside a rectangle that measured 9 cm × 9 cm, and participants sat approximately 56 cm from the computer monitor. For the no cue (NC) conditions, a fixation cross was presented at the center of the display. For the Cue groups, the fixation cross was replaced by a cue.

Procedure

The participants were randomly assigned to one of four groups. The two NC groups provided a replication of Lupyan and Spivey (2008). They were presented with stimulus arrays (**Figure 1A**) and instructed to identify whether the display was composed of a homogenous set of characters, or whether the display included one character that was different than the others. One of the NC groups was told that the display contained 2's and 5's whereas the other NC group was told that the displays contained abstract forms. For the two Cue groups, the fixation point was replaced with a visual cue (**Figure 1B**). For these participants, the task was to determine if an array item matched the cue. As with the NC conditions, one of the Cue groups was told that the display consisted of 2's and 5's and the other Cue group was told that the display contained abstract forms.

Each trial started with the onset of either a fixation cross (NC groups) or cue (CUE groups). The search array was added to the display after a 300-ms delay. Participants responded on one of two keys, indicating if the display contained one item that was different than the other display items. Following the response, an accuracy feedback screen was presented on the monitor for 1000 ms. The screen was then blanked for a 500-ms inter-trial interval. Average RT and accuracy were displayed at the end of each block.

The experiment consisted of a practice block of 12 trials and four test blocks of 60 trials each. At the beginning of each block, participants in both the NC and Cue groups were informed which character would be the target for that block of trials, similar to the procedure used by Lupyan and Spivey (2008). Each character served as the oddball for two of the blocks. The oddball was present on 50% of trials, positioned on the right and left side of the screen with equal frequency.

At the end of the experiment, the participants completed a short questionnaire to assess their strategy in performing the task. We were particularly interested in identifying participants in the abstract groups who had generated verbal labels for the rotated 2's and 5's given that such strategies produced a similar pattern of results as the Cue group in the Lupyan and Spivey (2008) study. Three participants in the NC group and two participants in the Cue reported using verbal labels, either spontaneously recognizing that the symbols were tilted 2's and 5's, or creating idiosyncratic labels (one subject reported labeling the items "valleys" and "mountains"). These participants were replaced, yielding a total of 12 participants in each of the four groups for the analyses reported below.

RESULTS

Overall, participants were correct on 89% of the trials and there was no indication of a speed–accuracy trade-off. Excluding incorrect trials, we analyzed the RT data (**Figure 2**) in a three-way ANOVA with two between-subject factors (1) task description (linguistic vs. abstract) and (2) task set (NC vs. Cue), and one within-subject factor, (3) set size (4, 6, or 10 items). The effect of set size was highly reliable, consistent with a serial search process, $F(2, 88) = 289.35$, $p < 0.0001$. Importantly, the two-way interaction of task description and task set was reliable, $F(1, 44) = 4.96$, $p < 0.05$, and there was also a significant three-way interaction, $F(2, 88) = 6.23$, $p < 0.005$, reflecting the fact that the linguistic advantage was greatest for the largest set size, but only for the NC group.

To explore these higher-order interactions, we performed separate analyses on the NC and Cue groups. For the NC groups, the data replicate the results reported in Lupyan and Spivey (2008). Participants who were instructed to view the characters as rotated numbers (linguistic description) responded much faster compared to participants for whom the characters were described as abstract symbols. Overall, the RT advantage was 303 ms, $F(1, 22) = 10.12$, $p < 0.001$.

We used linear regression to calculate the slope of the search functions, restricting this analysis to the target present data. The mean slopes for the linguistic and symbol groups were 112 and 143 ms, respectively. This difference was not reliable, ($p = 0.10$). However, there was one participant in the symbol group with a negative slope (−2 ms/item), whereas the smallest value for all of the other participants in this group was at least 93 ms/item. When the analysis was repeated without this participant, the mean slope for the symbol group rose to 155 ms/item, a value that was significantly higher than for the linguistic group ($p = 0.03$). In summary, consistent with Lupyan and Spivey (2008), the linguistic cues not only led to faster RTs overall, but also yielded a more efficient visual search process.

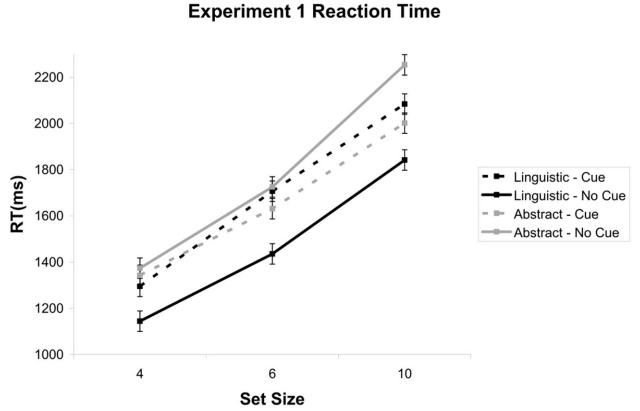


FIGURE 2 | Reaction time data for Experiment 1, combined over target present and target absent trials. Confidence intervals in the figure were calculated using the three-way interaction (Loftus and Masson, 1994).

A very different pattern of results was observed in the analysis of the data from the two Cue groups. Here, the linguistic advantage was completely abolished. In fact, mean RTs were slower by 46 ms for participants who were instructed to view the characters as rotated numbers, although this difference was not reliable $F(1, 22) = 0.072$, ns. Similarly, there was no difference in the efficiency of visual search, with mean slopes of 126 and 105 ms/item for the linguistic and symbol conditions, respectively. Thus, when the demands on working memory were reduced by the inclusion of a cue, we observed no linguistic benefit.

The results of Experiment 1 challenge the hypothesis that linguistic labels provide a top-down priming input to perceptual feature detectors. If this were so, then we would expect to observe a linguistic advantage regardless of whether the task involved a standard visual search (oddball detection) or our modified, matching task. *A priori*, we would expect that with either display, the linguistic description of the characters should provide a similar priming signal.

In contrast, the results are consistent with our working memory account. In particular, we assume that the linguistic advantage in the NC condition arises from the fact that participants must compare items in working memory during serial search, and that this process is more efficient when the display items can be verbally coded. Mean reaction time was faster and search more efficient (e.g., lower slope) when the rotated letters were associated with verbal labels. In this condition, each item can be assessed to determine if it matches the designated target, with the memory of the target facilitated by its verbal label (especially relevant here given that each target was tested in separate blocks). When the rotated letters were perceived as abstract symbols, the comparison process is slower, either because there is no verbal code to supplement the working memory representation of the target, or because participants end up making multiple comparisons between the different items.

The linguistic advantage was abolished when the target was always presented as a visual cue in the display. We can envision two ways in which the cue may have altered performance on the task. First, it would reduce the demands on working memory given that the cue provides a visible prompt. Second, it eliminates the need for comparisons between items in the display since each item can be successively compared to the cue. By either or both of these hypotheses, we would not expect a substantive benefit from verbal labels. RTs increase with display size, but at a similar rate for the linguistic and abstract conditions.

Mean RTs were slower for the Cue group compared to the NC group when the targets were described linguistically. This result

might indicate that the inclusion of the cues introduced some sort of interference with the search process. However, this hypothesis fails to account for why the slower RTs in the Cue condition were only observed in the linguistic group; indeed, mean RT was faster in the Cue condition for the abstract group. One would have to posit a rather complex model in which the inclusion of the cue somehow negated the beneficial priming from verbal labels.

Alternatively, the inclusion of the cue can be viewed as changing the search process in a fundamental way, with the task now more akin to a physical matching task rather than a comparison to a target stored in working memory. *A priori*, we cannot say which process would lead to faster RTs. However, the comparison of the absolute RT values between the Cue and NC conditions is problematic given the differences in the displays. One could imagine that there is some general cost associated with orienting to the visual cue at the onset of the displays for the Cue groups. Nonetheless, if the verbal labels were directly influencing perceptual detectors, we would have expected to see a persistent verbal advantage in the Cue condition, despite the slower RTs. The absence of such an advantage underscores our main point that the performance changes in visual search for the NC condition need not reflect differences in perception *per se*.

EXPERIMENT 2

We take a different approach in Experiment 2, testing the prediction that linguistic labels can disrupt processing when this information is task irrelevant. To this end, we had participants make an oddball judgment based on a physical attribute, line thickness. We presented upright or rotated 2s and 5s, assuming that upright numbers would be encoded as linguistic symbols, while rotated numbers would not. If language enhances perception, performance should be better for the upright displays. Alternatively, the automatic activation of linguistic codes for the upright displays may produce response conflict given that this information is irrelevant to the task.

MATERIALS AND METHODS

Participants

Twelve participants received class credit for completing the study.

Stimuli

Thick and thin versions of each character were created. The thick version was the same as in Experiment 1. For the thin version, the stroke thickness of each character was halved.

Procedure

Each trial began with the onset of a fixation cross for 300 ms. An array of four characters was then added to the display and remained visible for 450 ms (**Figure 3**). Participants were instructed to indicate whether the four characters had the same thickness, or whether one was different. The characters were either displayed in an upright orientation or rotated, with the same orientation used for all four items in a given display. Upright and rotated trials were randomized within a block. Each participant completed four blocks of 80 trials each. All other aspects were identical to Experiment 1.

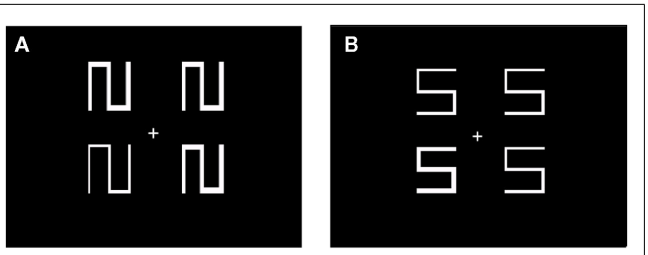


FIGURE 3 | Sample stimulus displays for Experiment 2. The four display items were letters, rotated 90°clockwise **(A)** or upright **(B)**. Participants made speeded responses, indicating if the font thickness for the displays items were homogenous or contained an oddball.

RESULTS

Participants were slower when the characters were upright compared to when they were rotated, $F(1, 11) = 7.67$, $p < 0.01$. The mean RT was 375 ms for the upright displays and 348 ms for the rotated displays, for an average cost of 27 ms ($SE_{diff} = 5.6$ ms). Participants averaged 92% correct, and there was no evidence of a speed accuracy trade-off.

We designed this experiment under the assumption that the upright displays would produce automatic and rapid activation of the lexical codes associated with the numbers, and that these task-irrelevant representations would disrupt performance on the thickness judgments. We can envision at least two distinct ways in which linguistic codes might disrupt performance. Perceptually, linguistic encoding encourages holistic processing. If parts of a number are thick, there is a tendency to treat the shape in a homogenous manner, perhaps reflecting the operation of categorization (Fuchs, 1923; Prinzmetal and Keysar, 1989; Khurana, 1998). This bias may be reduced for the less familiar, rotated shapes, which may be perceived as separate lines.

Alternatively, the linguistic codes could provide potentially disruptive input to decision processes (e.g., response selection). This hypothesis is similar to the theoretical interpretation of the Stroop effect (MacLeod, 1991). In the classic version of that task, interference is assumed to arise from the automatic activation of the lexical codes of word names when the task requires judging the stimulus color, at least when both the relevant and irrelevant dimensions map onto similar response codes (e.g., verbal responses). In the current task, this interference would be more at a conceptual level (Ivry and Schlerf, 2008). Given that the four items in the display were homogenous, we would expect priming of the concept "same", relative to the concept "different", and that this would occur more readily for the upright condition where the items are readily recognized as familiar objects.

DISCUSSION

In the current study, we set out to sharpen the focus on how language influences perception. This question has generated considerable interest, reflecting the potential utility for theories of embodied cognition to provide novel perspectives on the psychological and neural underpinnings of abstract thought (Gallese and Lakoff, 2005; Feldman, 2006; Barsalou, 2008). An explosion of

empirical studies have appeared, providing a wide range of intriguing demonstrations of how behavior (reviewed in Barsalou, 2008) and physiology (Thierry et al., 2009; Landau et al., 2010; Mo et al., 2011) in perceptual tasks can be influenced by language. We set out here to consider different ways in which language might influence perceptual performance.

As a starting point, we chose to revisit a study in which performance on a visual search task was found to be markedly improved when participants were instructed to view the search items as linguistic entities, compared to when the instructions led the participants to view the items as abstract shapes (Lupyan and Spivey, 2008). The authors of that study had championed an interpretation and provided a computational model in which over-learned associative links between linguistic and perceptual representations allowed top-down effects of a linguistic cue to sharpen perceptual analysis.

While this idea is certainly plausible, we considered an alternative hypothesis, one that shifts the focus away from a linguistic modulation of perceptual processes. In particular, we asked if the benefit of the linguistic cues might arise because language, as a ready form of efficient coding, might reduce the burden on working memory. We tested this hypothesis by using identical search displays, with the one addition of a visual cue, assumed to minimize the demands on working memory. Under these conditions, we failed to observe any performance differences between participants given linguistic and non-linguistic prompts. These results present a challenge for the perceptual account, given the assumption that top-down priming effects would be operative for both the cued and non-cued versions of the task. Instead, the working memory hypothesis provides a more parsimonious account of the results, pointing to subtle ways in which performance entails a host of complex operations.

Our emphasis on how language might influence performance at post-perceptual stages of processing is in accord with the results from studies employing a range of tasks. In a particularly clever study, Mitterer et al. (2009) showed that linguistic labels bias the reported color of familiar objects. When presented with a picture of a standard traffic light in varying hues ranging from yellow to orange, German speakers were more likely to report the color as "yellow" compared to Dutch speakers, a bias consistent with the labels used by each linguistic group. Given the absence of differences between the two groups in performance with neutral stimuli, the authors propose that the effect of language is on decision processes, rather than by directly influencing perception.

It should be noted, however, that participants in the Mitterer et al. (2009) study were not required to make speeded responses; as such, this study may be more subject to linguistic influences at decision stages than would be expected in a visual search task. However, numerous visual search studies have also shown that RT in such studies is influenced by the degree and manner in which targets and distractors are verbalized (Jonides and Gleitman, 1972; Reicher et al., 1976; Wang et al., 1994). Consistent with the current findings, RTs are consistently slower when the stimuli are unfamiliar, an effect that has been attributed to the more efficient processing within working memory for familiar, nameable objects (e.g., Wang et al., 1994).

We recognize that language may have an influence at multiple levels of processing. That is, the perceptual and working memory accounts are not mutually exclusive, and in fact, divisions such as "perception" and "working memory" may in themselves be problematic given the dynamics of the brain. Nonetheless, we do think there is value in such distinctions since it is easy for our descriptions of task domains to constrain how we think about the underlying processes.

Indeed, this concern is relevant to some work conducted in our own lab. In a series of studies, we have shown that the effects of language on visual search is more pronounced in the right visual field (Gilbert et al., 2006, 2008). We have used a simple visual search task here, motivated by the goal of minimizing demands on memory processes and strategies. Our results, showing that task-irrelevant linguistic categories influence color discrimination, can be interpreted as showing that language has selectively shaped perceptual systems in the left hemisphere. Alternatively, activation of (left hemisphere) linguistic representations may be retrieved more readily for stimuli in the right, compared to left, visual field, and thus exert a stronger influence on performance. While the answer to this question remains unclear – and again, both hypotheses may be correct – the visual field difference disappears when participants perform a concurrent verbal task (Gilbert et al., 2006, 2008). This dual-task result provides perhaps the most compelling argument against a linguistically modified structural asymmetry in the perceptual systems of the two hemispheres. Rather, it is consistent with the post-perceptual account promoted here (see also Mitterer et al., 2009) given the assumption that the secondary task disrupted the access of verbal codes for the color stimuli, an effect that would be particular pronounced in the left hemisphere.

We extended the basic logic of our color studies in the second experiment presented here, designing a task in which language might hinder perceptual performance. We again used a visual search task, but one in which participants had to determine if a display item had a unique physical feature (i.e., font thickness). For this task, linguistic representations were irrelevant. Nonetheless, when the shapes were oriented to facilitate reading, a cost in RT was observed, presumably due to the automatic activation of irrelevant linguistic representations.

While linguistic coding can be a useful tool to aid processing, the current findings demonstrate that language can both facilitate and impede performance. Language can provide a concise way to categorize familiar stimuli; in visual search, linguistic coding would provide an efficient mechanism to encode and compare the display items (Reicher et al., 1976; Wang et al., 1994). However, when the linguistic nature of the stimulus is irrelevant to the task, language may also hurt performance (Brandimonte et al., 1992; Lupyan et al., 2010).

These findings provide a cautionary note when we consider how language and perception interact. No doubt, the words we speak simultaneously reinforce and compete with the dynamic world we perceive and experience. When language alters perceptual performance, is it tempting to infer a shared representational status of linguistic and sensory representations. However, even performance in visual search reflects memory, decision, and perceptual processes. We must be vigilant in characterizing the manner in which language and perception interact.

REFERENCES

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645.

Bartlett, J. C., Till, R. E., and Levy, J. C. (1980). Retrieval characteristics of complex pictures: effects of verbal encoding. *J. Verb. Learn. Verb. Behav.* 19, 430–449.

Borghi, A. M., and Pecher, D. (2011). Introduction to the special topic embodied and grounded cognition. *Front. Psychol.* 2:187. doi:10.3389/fpsyg.2011.00187

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cogn. Psychol.* 43, 1–22.

Brandimonte, M. A., Hitch, G. J., and Bishop, V. M. (1992). Verbal recoding of visual stimuli impairs mental image transformation. *Mem. Cognit.* 20, 449–455.

Ervin-Tripp, S. (1967). An Issei learns English. *J. Soc. Issues* 2, 78–90.

Feldman, J. (2006). From molecule to metaphor: A natural theory of language. Cambridge, MA: MIT Press.

Fuchs, W. (1923). "The influence of form on the assimilation of colours," in *A Source Book of Gestalt Psychology,* ed. W. D. Ellis (London: Routledge & Kegan Paul), 95–103.

Gallese, V., and Lakoff, G. (2005). The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. *Cogn. Neuropsychol.* 22, 455–479.

Gilbert, A. L., Regier, T., Kay, P., and Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Brain Lang.* 103, 489–494.

Gilbert, A. L., Regier, T., Kay, P., and Ivry, R. B. (2008). Support for lateralization of the Whorf effect beyond the realm of color discrimination. *Brain Lang.* 105, 91–98.

Goldstone, R. L., and Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition* 65, 231–262.

Ivry, R. B., and Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends Cogn. Sci. (Regul. Ed.)* 12, 273–280.

Jonides, J., and Gleitman, H. (1972). A conceptual category effect in visual search: O as letter or as digit. *Percept. Psychophys.* 12, 457–460.

Kay, P., and Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *Am. Anthropol.* 86, 65–79.

Khurana, B. (1998). Visual structure and the integration of form and color. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1766–1785.

Landau, A. N., Aziz-Zadeh, L., and Ivry, R. B. (2010). The influence of language on perception: listening to sentences about faces affects the perception of faces. *J. Neurosci.* 30, 15254–15261.

Loftus, G. R., and Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychon. Bull. Rev.* 1, 476–490.

Loftus, E. F., and Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *J. Verb. Learn. Verb. Behav.* 13, 585–589.

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42.

Lupyan, G. (2008). The conceptual grouping effect: categories matter (and named categories matter more). *Cognition* 108, 566–577.

Lupyan, G., and Spivey, M. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli. *Curr. Biol.* 18, R410–R412.

Lupyan, G., and Spivey, M. (2010). Making the invisible visible: auditory cues facilitate visual object detection. *PLoS ONE* 5, e11452. doi:10.1371/journal.pone.0011452

Lupyan, G., Thompson-Schill, S. L., and Swingley, D. (2010). Conceptual penetration of visual processing. *Psychol. Sci.* 21, 682–691.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychol. Bull.* 109, 163–203.

Mazer, J. A., and Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron* 40, 1241–1250.

Meteyard, L., Bahrami, B., and Vigliocco, G. (2007). Motion detection and motion verbs: language affects low-level visual perception. *Psychol. Sci.* 18, 1007–1013.

Mitterer, H., Horschig, J. M., Musseler, J., and Majid, A. (2009). The influence of memory on perception: it's not what things look like, it's what you call them. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1557–1562.

Mo, L., Xu, G., Kay, P., and Tan, L. H. (2011). Electrophysiological evidence for the left-lateralized effect of language on preattentive categorical perception of color. *Proc. Nat. Acad. Sci. U.S.A.* 108, 14026–14030.

Paivio, A. (1971). *Imagery and Verbal Processes.* New York: Holt, Rinehart & Winston.

Pinker, S. (1997). *How the Mind Works.* New York: W.W. Norton.

Prinzmetal, W., and Keysar, B. (1989). A functional theory of illusory conjunctions and neon colors. *J. Exp. Psychol. Gen.* 118, 165–190.

Reicher, G. M., Snyder, C. R. R., and Richards, J. T. (1976). Familiarity of background characters in visual scanning. *J. Exp. Psychol. Hum. Percep. Perform.* 2, 522–530.

Rosch, E. H. (1973). Natural categories. *Cogn. Psychol.* 4, 328–350.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., and Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proc. Nat. Acad. Sci. U.S.A.* 106, 4567–4570.

Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136.

Wang, Q., Cavanagh, P., and Green, M. (1994). Familiarity and pop-out in visual search. *Percept. Psychophys.* 56, 495–500.

Whorf, B. L. (1956). "Science and linguistics," in *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf,* ed. J. B. Carroll (Cambridge, MA: MIT Press), 207–219.

Wolfe, J. M. (1992). The parallel guidance of visual attention. *Curr. Dir. Psychol. Res.* 1, 124–128.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 October 2011; accepted: 01 March 2012; published online: 20 March 2012.

Citation: Klemfuss N, Prinzmetal W and Ivry RB (2012) How does language change perception: a cautionary note. Front. Psychology ***3****:78. doi: 10.3389/fpsyg.2012.00078*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 September 2011
doi: 10.3389/fpsyg.2011.00227

Abstract and concrete sentences, embodiment, and languages

Claudia Scorolli[1]*, Ferdinand Binkofski[2], Giovanni Buccino[3], Roberto Nicoletti[4], Lucia Riggio[5] and Anna Maria Borghi[1,6]*

[1] *Department of Psychology, University of Bologna, Bologna, Italy*
[2] *Division for Cognitive Neurology, RWTH Aachen, Aachen, Germany*
[3] *Department of Medical Sciences, University of Catanzaro, Catanzaro, Italy*
[4] *Department of Communication Disciplines, University of Bologna, Bologna, Italy*
[5] *Department of Neurosciences, University of Parma, Parma, Italy*
[6] *Institute of Cognitive Science and Technology, CNR, Rome, Italy*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Klaus Kessler, University of Glasgow, UK
Lawrence Taylor, Northumbria University, UK

****Correspondence:***
Claudia Scorolli, Department of Psychology, University of Bologna, Viale Berti Pichat, 5-40127 Bologna, Italy.
e-mail: claudia.scorolli2@unibo.it;
Anna Maria Borghi, Department of Psychology, University of Bologna, Viale Berti Pichat, 5-40127 Bologna, Italy.
e-mail: annamaria.borghi@unibo.it

One of the main challenges of embodied theories is accounting for meanings of abstract words. The most common explanation is that abstract words, like concrete ones, are grounded in perception and action systems. According to other explanations, abstract words, differently from concrete ones, would activate situations and introspection; alternatively, they would be represented through metaphoric mapping. However, evidence provided so far pertains to specific domains. To be able to account for abstract words in their variety we argue it is necessary to take into account not only the fact that language is grounded in the sensorimotor system, but also that language represents a linguistic–social experience. To study abstractness as a continuum we combined a concrete (C) verb with both a concrete and an abstract (A) noun; and an abstract verb with the same nouns previously used (grasp vs. describe a flower vs. a concept). To disambiguate between the semantic meaning and the grammatical class of the words, we focused on two syntactically different languages: German and Italian. Compatible combinations (CC, AA) were processed faster than mixed ones (CA, AC). This is in line with the idea that abstract and concrete words are processed preferentially in parallel systems – abstract in the language system and concrete more in the motor system, thus costs of processing within one system are the lowest. This parallel processing takes place most probably within different anatomically predefined routes. With mixed combinations, when the concrete word preceded the abstract one (CA), participants were faster, regardless of the grammatical class and the spoken language. This is probably due to the peculiar mode of acquisition of abstract words, as they are acquired more linguistically than perceptually. Results confirm embodied theories which assign a crucial role to both perception–action and linguistic experience for abstract words.

Keywords: abstract concepts, embodiment, social–linguistic experience, cross-language comparison, parallel processing

INTRODUCTION

The distinction between "abstract" and "concrete" concepts and words is all but uncontroversial. People disagree when trying to categorize a specific noun as "abstract," and even more when classifying as such a specific verb. Evidence suggests that the "abstract–concrete dimension" reflects a continuum rather than a dichotomy. Indeed, Nelson and Schreiber (1992) and Wiemer-Hastings et al. (2001) asked people to judge the concreteness of large sets of words; they found a bimodal distribution (according to features, such as tangibility or visibility), not a dichotomy. Things are even more complicated when words are embedded within contexts. Most of us would agree that the noun "apple" and the verb "to grasp" are concrete, but judging verb–noun pairs such as "to grasp the meaning," or "to think about an apple" (e.g., Aziz-Zadeh et al., 2006) is all but simple. In addition, the meaning of a sentence is often influenced by a specific language and culture; furthermore, it has been shown that this linguistic and cultural influence is particularly strong for abstract compared to concrete words (Boroditsky, 2003).

The study of how abstract concepts and words are represented has been the focus of many investigations in the 1960s–1990s. The two most influential views were the context availability theory (CAT, Schwanenflugel, 1991) and the dual coding theory (DCT, e.g., Paivio, 1986). CAT would ascribe the processing difference between concrete and abstract words to the fact that concrete words have stronger semantic relations with the context represented by other words. According to DCT, instead, abstract words would be represented only in a linguistic system while concrete words would be represented both in imagery and linguistic system.

As to the neural substrates of language comprehension, the integration of lesions analyses, white matter tractography, and resting state functional magnetic resonance imaging (e.g., Dronkers et al.,

2004; Turken and Dronkers, 2011) have recently brought into question traditional models: not only the left posterior temporal cortex but an extensive network in the left hemisphere seems to be critical for the processing of language (e.g., left posterior middle temporal gyrus, MTG; the anterior part of Brodmann's area 22; the posterior superior temporal sulcus). The investigation of the structural and functional connectivity of the keys regions (using diffusion tensor imaging) has shown a *bilateral* temporo-parieto-frontal network supported by long-distance white matter pathways. This network seems to interact with other brain regions outside the traditionally recognized language areas (Turken and Dronkers, 2011). Pertaining to the aim of the present work, in the last years we have assisted a renewed interest for the way concrete and abstract words are represented, as the growing body of brain imaging studies reveals (e.g., Desai et al., 2010; Ghio and Tettamanti, 2010). Many of these studies supported the original proposal by Paivio, showing for example that processing of abstract words is more lateralized in the left hemisphere than processing of concrete ones (for a review see Binder et al., 2005).

In the same line, on the theoretical side it has been recently proposed that language comprehension is both embodied and symbolic (e.g., Louwerse and Jeauniaux, 2008; Dove, 2010). In keeping with Paivio, Dove (2009, 2010) argues in favor of "representational pluralism," claiming that perceptual simulations play an important role in highly imageable concepts while amodal linguistic representations play a crucial role in abstract concepts.

One of the reasons of the renewed interest for abstract words is that understanding the way we represent abstract words is a testbed for the increasingly popular (e.g., Chatterjee, 2010) embodied theories of language comprehension, according to which language is grounded in perception, action, and emotional system (for reviews, see Barsalou, 2008; Fischer and Zwaan, 2008; Gallese, 2008). Whereas it is now widely recognized that the evidence in support of embodied theories is compelling regarding concrete or highly imageable words, the issue is much debated regarding abstract words and sentences (Pezzulo and Castelfranchi, 2007; Louwerse and Jeauniaux, 2008; Dove, 2010). Within the embodied framework abstract words would be explained as the result of the transfer in abstract domains of image-schemas derived from sensorimotor experiences: for example, the image-schema derived from "container" would be used to understand the notion of "category" (Lakoff, 1987; Gibbs and Steen, 1999; Boot and Pecher, 2011), the action of giving a concrete object (pizza) would be used to understand the action of giving some news (Glenberg et al., 2008). Alternatively, it has been proposed that abstract words evoke different kinds of properties, i.e., that they activate situations and introspective relationships more frequently than concrete words (Barsalou, 1999; Barsalou and Wiemer-Hastings, 2005; for a review see Pecher et al., 2011).

More crucial to our work are some recent proposals which, starting from an embodied perspective and avoiding assuming the existence of amodal symbols, detached from perceptual and motor experience, share with Paivio the idea that multiple types of representation underlie knowledge (for a review see special topic on Embodied and Grounded Cognition, Borghi and Pecher, 2011). These proposals differ from Paivio's view as they hypothesize that not only concrete, but also abstract words are embodied and grounded. According to the language and situated simulation (LASS) theory (Barsalou et al., 2008), linguistic forms and situated simulations interact continuously and different mixtures of the two systems underlie a wide variety of tasks. The linguistic system (comprising the left-hemisphere language areas, and especially the left inferior frontal gyrus, Broca's area) is involved mainly during superficial linguistic processing, whereas a deeper conceptual processing necessarily requires the simulation system, made up of the bilateral posterior areas associated with mental imagery and episodic memory.

The word as social tools (WAT) proposal (Borghi and Cimatti, 2009) differs from the LASS theory because, according to WAT, the linguistic system does not simply involve a form of superficial processing: words are not conceived of as mere signals of something but also as *tools* that allow us to operate in the world. In addition, WAT extends LASS as it formulates more detailed predictions on the representation of abstract and concrete words. Indeed, according to WAT abstract word meanings would rely more than concrete word meanings on the everyday experience of being exposed to language in social contexts. According to WAT the difference between abstract and concrete words basically relies on the different *mode of acquisition* (MoA; Wauters et al., 2003), which can be perceptual, linguistic, or mixed. MoA ratings, which correlate but are not totally explained by age of acquisition, concreteness, and imageability, gradually change over grades. In the first grades acquisition is mainly perceptual, later it is mainly linguistic. It can follow that abstract words are typically acquired later, also because it is more difficult to linguistically explain a word meaning than to point at its referent while labeling. The acquisition of abstract words, due to their complexity, typically require a long-lasting social interaction, and it often implies complex linguistic explanations and repetitions. In contrast, the process by which young children learn concrete words appears effortless and often occurs within a single episode of hearing the word spoken in context (e.g., Carey, 1978; see also Pulvermüller, in press). This has the consequence that, even if for the representation of both concrete and abstract words meanings sensorimotor and linguistic experience are crucial, we rely more on language to understand the meaning of concrete words, whereas we rely more on non-linguistic sensorimotor experience to grasp the meaning of abstract words. (Borghi and Cimatti, 2009). Given that abstract words do not have a specific object or entity as referent, many of them might be acquired linguistically, i.e., listening to other people explaining their content to us, rather than perceptually. This might be due also to their different degree of complexity: learning to use a word such as "lipstick" is simpler than learning to use a word like "justice," and the linguistic label might be more crucial for keeping together experiences as diverse as those related to the notion of "justice." Borghi et al. (2011) used novel categories to mimic the acquisition of concrete and abstract concepts; they found that linguistic explanations are more important for the acquisition of abstract than for concrete words, and showed with a property verification task that concrete words evoke more manual information, while abstract words elicit more verbal information. WAT hypothesizes also that the MoA determines the representation of the word in our brain: when the words refer to categories learned through sensorimotor experiences (e.g., "bottle"), they have a much higher level of grounding

in the perception and action systems than words learned mainly through the mediation of other words (e.g., "democracy"; Borghi et al., 2011; see also Prinz, 2002). Consistently, concrete words should evoke more manual information, activating precociously motor areas (Jirak et al., 2010; Pulvermüller, in press), whereas abstract words should elicit more verbal-linguistic information, activating precociously motor areas related to the mouth, as data on transcranial magnetic stimulation study (Scorolli et al., 2011) and on words acquisition modality suggest (Borghi et al., 2011).

Notice that claiming that concrete and abstract words are acquired through different modalities does not require the postulation of any difference in format between the two kinds of words, nor any transduction from sensorimotor experience into amodal symbols. It simply means that abstract word meanings should rely more on the embodied experience of being exposed to language than concrete word meanings. However, we do not intend to imply that abstract words rely on the simple embodied experience of speaking and listening – this would not suffice to call their representation embodied. In contrast with non-embodied approaches to abstract words, in our view a word like "philosophy" would activate perceptual and motor experiences, together with linguistic experience. As demonstrated by Borghi et al. (2011), with abstract terms the advantage of linguistic over manual information was present only when linguistic information did not contrast with perceptual one.

The major difference between Paivio's approach and multiple representation theories such as WAT's approach to concrete and abstract words is that, according to the first, abstract words rely only on the verbal system, while for WAT both concrete and abstract words are grounded in perception and action systems, even if the linguistic system plays a major role for abstract words representation.

The best way to disambiguate these hypotheses is the selection of a paradigm that allows contrasting abstract and concrete words combined in sentences. So far most evidence has been found with brain imaging rather than with behavioral studies, it concerns single words rather than words embedded in contexts, and tasks requiring deep semantic processing are typically not used [an exception is given by a recent fMRI study by Desai et al. (2010), in which a sentence evaluation task was used]. In contrast, our study focuses on how words meaning changes depending on the context in which it is embedded. For this reason we will compare not only *whole* abstract and concrete sentences, but also sentences which result from a mixture of abstract and concrete nouns and verbs in a well-balanced design. We believe this may represent an important step for a systematic investigation of abstraction. One of the advantages of this design resides in the possibility to study abstractness in a continuum, and to verify the effects on comprehension using different combinations and studying how the meaning of single words can change depending on the context. In addition, focusing on sentences instead than on single words offers the possibility to investigate linguistic processing in a more ecological way, and allows us detecting eventual influences of the different spoken languages.

In the present study we asked participants to judge the sensibility of sentences. We chose this task because it is established that it implies a deep semantic processing of the sentences (see also Turken and Dronkers, 2011). Coherently with previous literature, we defined as "concrete" only nouns that refer to manipulable objects and only verbs referring to manual actions (e.g., "a flower"/"to grasp"). We decided to define as "abstract" only nouns that do not refer to an object, rather to an entity that can neither be grasped nor touched, and only verbs that refer to an action[1] that cannot be performed with any part of the body, that is, an action that does not explicitly require any movement or any activation of the motor system (e.g., "a concept"/"to describe"). In addition, to investigate the specific effects of the specific language we use, we examined different combinations of nouns (abstract and concrete ones) and verbs (abstract and concrete ones), in two languages, German and Italian, which are syntactically different: in German the noun precedes the verb; in Italian it is the opposite.

There are several possible views:

1. No difference view: abstract and concrete concepts have the same core representations. According to the amodal theories their representations in the brain would be most probably in the language domain; according to the strictly modal view both concrete and abstract concepts would be represented in the perception and action system.
2. Non-embodied multiple representation view: concrete and abstract words have distinct representations: the first are represented in the sensorimotor system, abstract words in the language system. This view, proposed by Paivio (1986), is adopted by multiple representation views not adopting an embodied approach to abstract words, i.e., to views arguing that concrete and abstract words differ in format (e.g., Binder et al., 2005; Dove, 2010).
3. Embodied multiple representation view: abstract and concrete concepts are represented both in the language domains and in the perception and action systems. However, they are not represented in the same way in the two systems but there is a different distribution. Linguistic information should be more relevant for abstract words, perception, and action information for concrete ones. This is the view consistent with multiple representation theories adopting an embodied perspective, such as WAT and LASS.

In contrast with strictly amodal and strictly modal views (No difference views), both embodied and non-embodied multiple representation views predict costs in mixed combinations, when switching from one perceptual modality to another (Pecher et al., 2003). In addition, according to the WAT proposal mixed combinations should be differently modulated by the syntactical structure of the two different chosen languages. As the Age of Acquisition clearly affects performance in semantic tasks (Lewis, 1999; Brysbaert et al., 2000) and is correlated with the Modality of Acquisition, WAT predicts that in mixed conditions RTs should be slower when the abstract word precedes the concrete one, due to the fact that the former is acquired later and relies more on linguistic information than the second (Bloom, 2000; Colombo and Burani, 2002; Mestres-Missé et al., 2009).

[1]Action thought in a more general way, as to also include cognitive processes, or mental operations.

EXPERIMENTAL METHOD

PARTICIPANTS

Thirty-eight students from the University of Hamburg (group I) and 38 students from the University of Bologna (group II) took part in the study. All were native German speakers (group I) or native Italian speakers (group II), right-handed according to the Edinburgh Handedness Questionnaire (Oldfield, 1971), and all had normal or corrected-to-normal vision. They all gave their informed consent to the experimental procedure. Their ages ranged from 18 to 32 years old (German group: $M = 26.26$; SD = 3.64; Italian Group: $M = 24.61$; SD = 3.58). The study was approved by the local ethic committees.

MATERIALS

Materials consisted of word pairs (sentences) composed of a transitive verb and a concept noun. To study the dimension abstract–concrete in a continuum we contrasted two kinds of Verbs (Concrete vs. Abstract) with two kinds of Nouns (Concrete vs. Abstract). We defined Concrete Nouns as nouns referring to graspable objects, Concrete Verbs as verbs referring to hand actions, Abstract Nouns as nouns that do not refer to manipulable objects, and Abstract Verbs as verbs that do not refer to motor actions. Therefore we created 192 sentences – 48 quadruples – in the German language and 192 sentences – 48 quadruples – in the Italian language. Each quadruple was constructed by pairing a Concrete Verb (e.g., to grasp) both with a Concrete Noun (e.g., a flower) and an Abstract Noun (e.g., a concept); and by pairing an Abstract verb (e.g., to describe) with the previously used concrete and abstract nouns (e.g., to squeeze/find a sponge/friendship; to lift/receive a table/criticism; to caress/wait for a dog/idea; to bend/respect the menu/will; to paint/admire the frame/sunset; to write/look for the document/end; to carve out/wait for a newspaper/moment). We decided to use sentences with a very simple grammatical structure (a verb plus a noun) as it was not possible to develop more complex sentences with a similar grammatical structure that fulfilled the criteria of the quadruples. The majority of these sentences' meanings matched in both languages; a few of them slightly differed, as some pairs did not allow for a literal translation.

Due to the different syntax of the German and Italian languages, the German sentences were composed of a noun followed by a verb; the Italian ones were composed of a verb followed by a noun. We chose to compare these two languages as the specific differences in the syntactical structure allowed us to speculate on the different effects caused by a verb preceded by a noun (German sample) vs. a noun preceded by a verb (Italian sample).

To select 30 critical quadruples from the 48 ones, we asked 20 German students and 20 Italian students to judge how familiar each sentence sounded and with what degree of probability they would use each sentence. They were required to provide ratings on a continuous scale (Not familiar – Very Familiar; Not probably – Very probably), by making a cross on a line. We selected the quadruples with highest scores for both familiarity and probability of use, and, from these, we finally chose the quadruples with lower scores in the SDs. Thus we obtained 120 verb–noun pairs (balanced for familiarity and probability of use).

Due to the peculiarity of our linguistic materials, to further test if the 120 selected verb–noun pairs differed as far as the frequency of use is concerned, we checked on the research engine "Google" the frequency of each pair, by using quotations marks (Page et al., 1998; Griffiths et al., 2007; Sha, 2010). The frequencies were submitted to a 2 (kind of Noun: Concrete vs. Abstract) × 2 (kind of Verb: Concrete vs. Abstract) × 2 (Language: German vs. Italian) ANOVA. Crucially we did not find any significant effect. This further control on written frequency prevented us from accounting for possible differences on processing resting on different association degrees between words pairs composing German and Italian quadruples.

In addition to the 30 critical quadruples, we created 30 filler quadruples using the same criteria. We combined a concrete verb both with a concrete noun and with an abstract noun; and we combined an abstract verb with the same concrete noun and abstract noun, leading to nonsensical sentences (e.g., "to switch off the shoe"). Each quadruple was presented only once.

PROCEDURE

German and Italian participants were randomly assigned to one of two groups. Members of both groups were tested individually in a quiet library room. They sat on a comfortable chair in front of a computer screen and were instructed to look at a fixation cross that remained on the screen for 1000 ms. Then a sentence appeared on the screen for 2600 ms. The German sentences were composed of a determinative or non-determinative article plus a noun plus a verb (example for the concrete noun – concrete verb combination: "einen Kuchen anschneiden," to cut a cake), while the Italian sentences were composed of a verb plus a determinative or non-determinative article plus a noun (example for the concrete verb – concrete noun combination: "stringere una spugna," to squeeze a sponge).

The timer started operating when the sentence appeared on the screen. For each verb–noun pair, participants were instructed to press one key if the combination made sense, and to press another key if the combination did not make sense.

Participants in the first group (both German and Italian) were asked to respond "yes" with their left hand and "no" with their right hand; participants in the other group (both German and Italian) were required to do the opposite. All participants were informed that their response times (RT) would be recorded and were invited to respond as quickly as possible while still maintaining accuracy. Stimuli were presented in a random order. The 240 experimental trials were preceded by 8 training trials, in order to allow the participants to familiarize themselves with the procedure.

STATISTICAL ANALYSIS

In our analyses we considered only the sensible sentences. Participants were accurate in responding; no participant's responses included errors over 15%. To screen for outliers, scores 2 SDs higher or lower than the mean participant score were removed for each participant. Removed outliers accounted for 3.6% of response trials. The remaining RT and errors were submitted to a 2 (kind of Noun: Concrete vs. Abstract) × 2 (kind of Verb: Concrete vs. Abstract) × 2 (Mapping: yes-right/no-left vs. yes-left/no-right) × 2 [Language: German: noun (first), verb (second) vs. Italian: noun (second), verb (first)] mixed factor ANOVA, with Mapping and Language as between-participants variables.

We conducted the analyses with participants as a random factor. As the error analysis revealed that there was no speed–accuracy trade-off, we will discuss only the RT analysis

ASSESSMENT OF GERMAN AND ITALIAN PAIRS

Materials were controlled regarding a variety of dimensions. 30 students from the University of Hamburg and 30 students from the University of Bologna were asked to rate the ease or difficulty with which each pair evoked mental images (imageability: Low imagery rate – High Imagery rate) on a continuous scale (scores ranging from 0 to 100); how literally they would take each pair (literality: Literal – No Literal); whether and to what extent each pair elicited movement information (quantity of motion: Not much movement – Much movement). Finally 10 German students and 10 Italian students were asked to rate at which age approximately they had learned to use each pair (age of acquisition ratings). For each rating, we calculated the scores' averages and the scores' SDs for each condition.

Imageability

Both German and Italian participants judged the Concrete Verb – Concrete Noun pairs as the easiest to imagine (see **Figure 1**, Germans: $M = 69.10$; $SD = 12.76$; Italians: $M = 77.74$; $SD = 8.49$), followed by the Abstract Verb – Concrete Noun pairs (Germans: $M = 52.72$; $SD = 15.80$; Italians: $M = 51.33$; $SD = 18.65$), by the Concrete Verb – Abstract Noun pairs (Germans: $M = 48.53$; $SD = 12.92$; Italians: $M = 46.33$; $SD = 12.36$), and finally by Abstract Verb – Abstract Noun pairs (Germans: $M = 45.56$; $SD = 14.51$; Italians: $M = 44.88$; $SD = 15.23$). Results showed that German and Italian participants had the same pattern: the pair containing two concrete words was judged as the easiest to imagine. Moreover for both groups the noun was stronger than the verb in determining the imageability of the sentence.

Literality–metaphoricity

German participants rated the Abstract verb – Concrete noun pairs as the ones that they would take most literally (see **Figure 2**, $M = 18.89$; $SD = 13.72$), followed by the Concrete Verb – Concrete Noun pairs ($M = 20.22$; $SD = 18.12$), by the Abstract Verb – Abstract Noun pairs ($M = 31.23$; $SD = 19.59$), and finally by the Concrete Verb – Abstract Noun pairs ($M = 56.95$; $SD = 19.01$). Italian participants rated the Concrete Verb – Concrete Noun pairs as the sentences that they would take most literally ($M = 11.42$; $SD = 4.57$), followed by the Abstract Verb – Concrete Noun pairs ($M = 31.33$; $SD = 13.11$), by the Abstract Verb – Abstract Noun pairs ($M = 59.42$; $SD = 13.63$), and finally by Concrete Verb – Abstract Noun pairs ($M = 69.50$; $SD = 11.78$).

The sentences rated as more literal are the ones which contained a Concrete Verb plus a Concrete Noun for Italian participants and containing an Abstract Verb plus a Concrete Noun for German participants. Both groups judged the combination Concrete Verb – Abstract Noun as the most metaphorical one. It is worth noting that while the concrete noun meaning remains the same through the quadruples, the concrete verb meaning, as well as its concreteness/abstractness, changes through the quadruples, depending on the context: for example, the meaning of the verb "to grasp" is not the same in "grasping an apple" and in "grasping a concept" (Parisi, personal communication).

Quantity of motion

German participants rated the Concrete Verb – Concrete Noun pairs as the ones that elicited most movement information (see **Figure 3**, $M = 34.29$; $SD = 13.95$), followed by the Concrete Verb – Abstract Noun pairs ($M = 27.22$; $SD = 12.82$), by the Abstract Verb – Abstract Noun pairs ($M = 17.98$; $SD = 13.87$) and finally by Abstract Verb – Concrete Noun pairs ($M = 13.99$; $SD = 7.39$). Interestingly, the Italian participants' pattern was different, as they rated the Concrete Verb – Abstract Noun pairs as the ones that mainly elicited movement information ($M = 42.56$; $SD = 13.28$), followed by the Abstract Verb – Abstract Noun pairs ($M = 35.05$; $SD = 12.24$), by the Concrete Verb – Concrete Noun pairs ($M = 31.93$; $SD = 10.58$) and finally by the Abstract Verb – Concrete Noun pairs ($M = 21.56$; $SD = 11.25$).

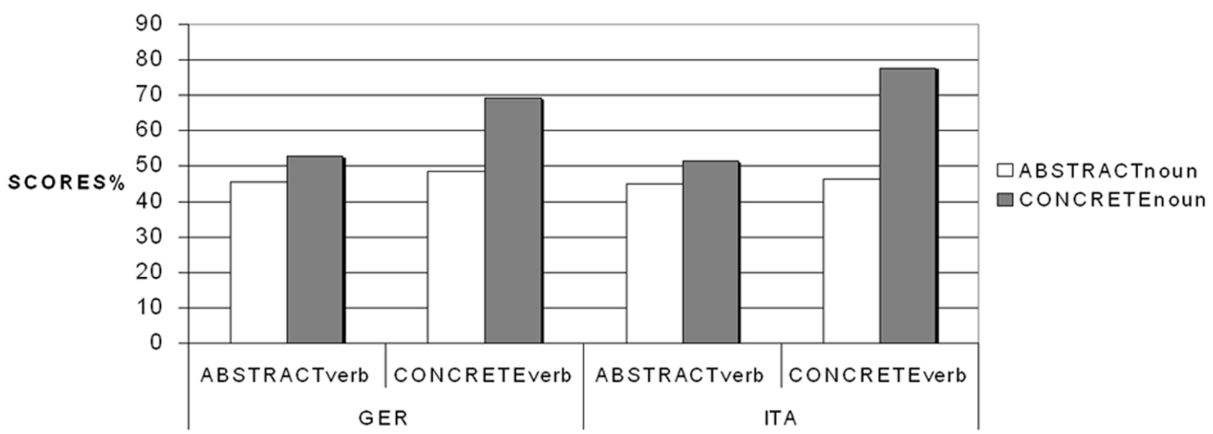


FIGURE 1 | German and Italian participants had the same pattern: the pair containing both words concrete was judged as the easiest to imagine. Moreover for both groups the noun was stronger than the verb in determining the imageability of the sentences.

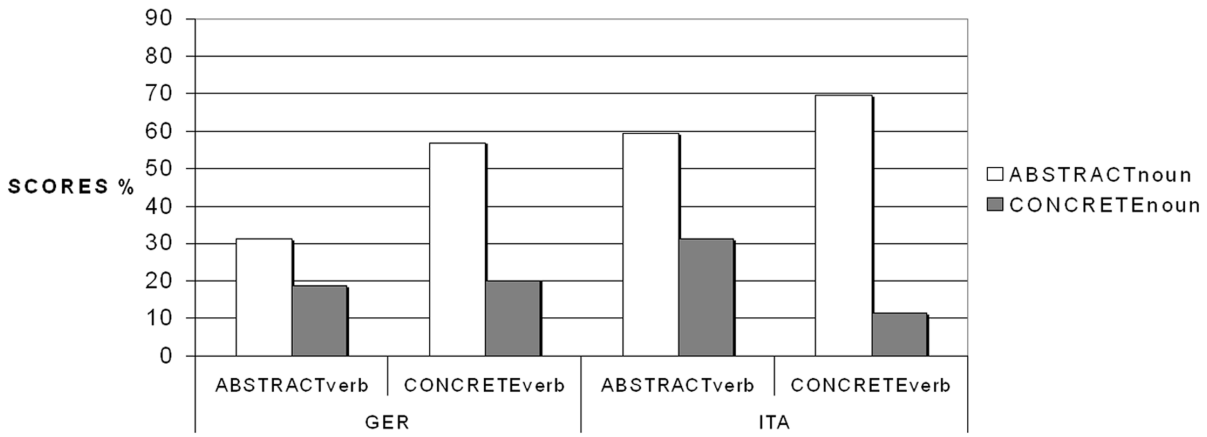


FIGURE 2 | Both groups judged the combination Concrete Verb plus Abstract Noun as the most metaphorical one. *Note:* while the concrete noun meaning remains the same through the quadruples, the concrete verb meaning, as well as its concreteness/abstractness, changes through the quadruples, depending on the context.

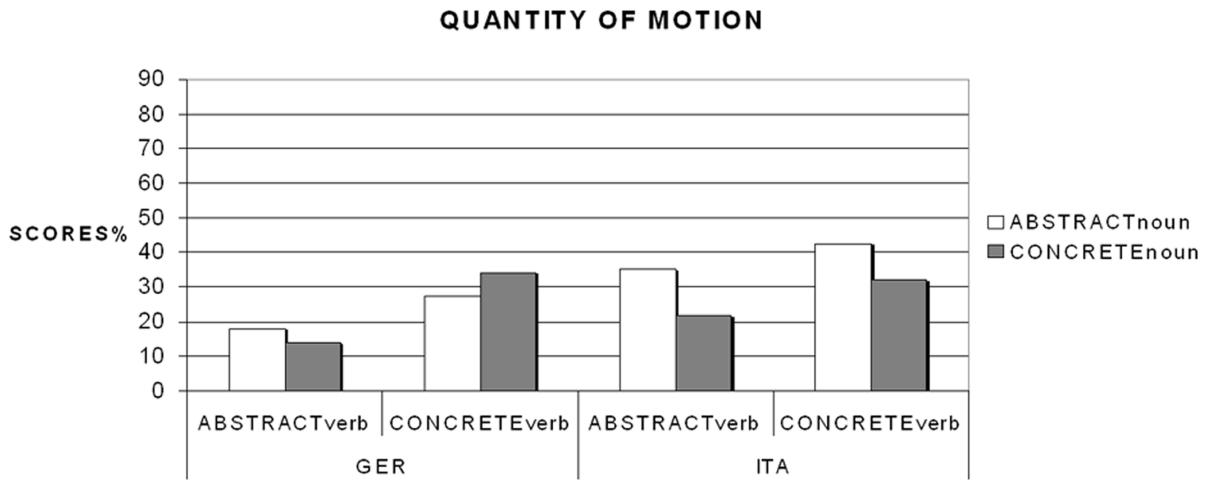


FIGURE 3 | Both groups agreed in judging the Abstract Verb plus Concrete Noun combination as the one that elicits less movement. The main difference concerns the Concrete Verb plus Abstract Noun *vs.* Concrete Verb plus Concrete Noun combinations: the former suggested the biggest amount of movement for Italian participants; the latter evoked the huger quantity of motion in German participants.

Both groups agreed in judging the Abstract Verb – Concrete Noun combination as the one that elicits less movement. The main difference concerns the combinations Concrete Verb – Abstract Noun vs. Concrete Verb – Concrete Noun combination, as while the former suggested the biggest amount of movement for Italian participants, the latter evoked the larger quantity of motion in German participants.

Age of acquisition

A number of studies (Gilhooly and Gilhooly, 1980; Zevin and Seidenberg, 2002) have demonstrated the validity of age of acquisition ratings, by showing that age rated by adults is the major independent predictor of the objective age of acquisition indices. In our study German participants rated the Concrete Verb – Concrete Noun pairs as the ones they learnt first (see **Figure 4**, $M = 7.82$ years old; SD $= 2.21$), followed by the Abstract Verb – Concrete Noun pairs ($M = 8.64$ years old; SD $= 2.55$), and finally by both Abstract Verb – Abstract Noun pairs and Concrete Verb – Abstract Noun pairs ($M = 10.24$ years old; SD $= 2.35$; $M = 10.74$ years old; SD $= 1.95$). The pattern was the same for Italian participants who rated the Concrete Verb – Concrete Noun pairs as the earliest learnt ones ($M = 6.63$ years old; SD $= 1.97$), followed by the Abstract Verb – Concrete Noun pairs ($M = 8.33$ years old; SD $= 2.34$), and finally by both Abstract Verb – Abstract Noun

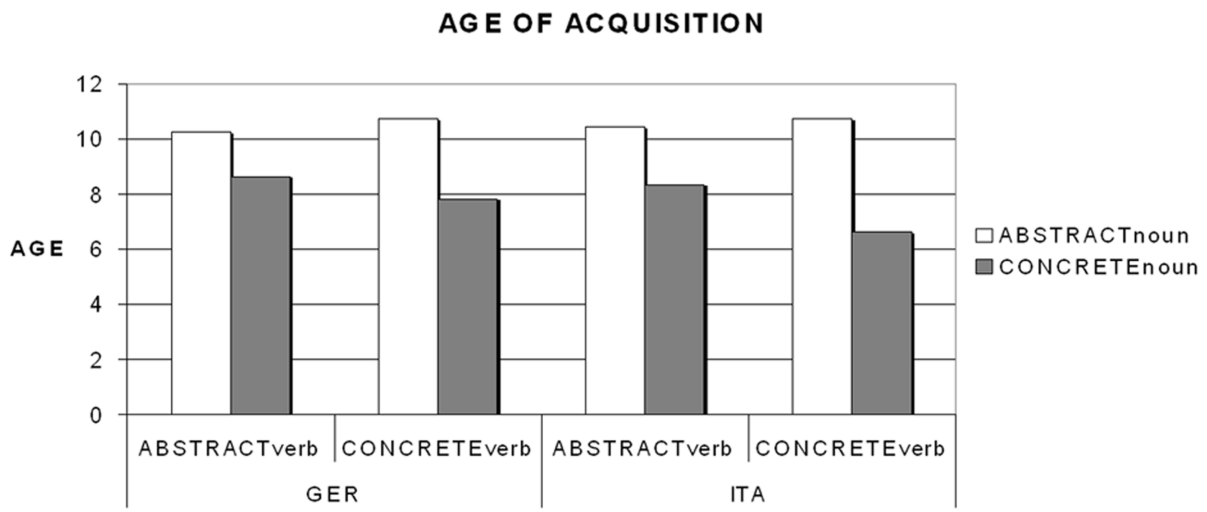


FIGURE 4 | Results suggest that the different age of acquisition of sentences is explained by the noun: as shown in the literature on single word age of acquisition, the concrete noun is learned before than the abstract one.

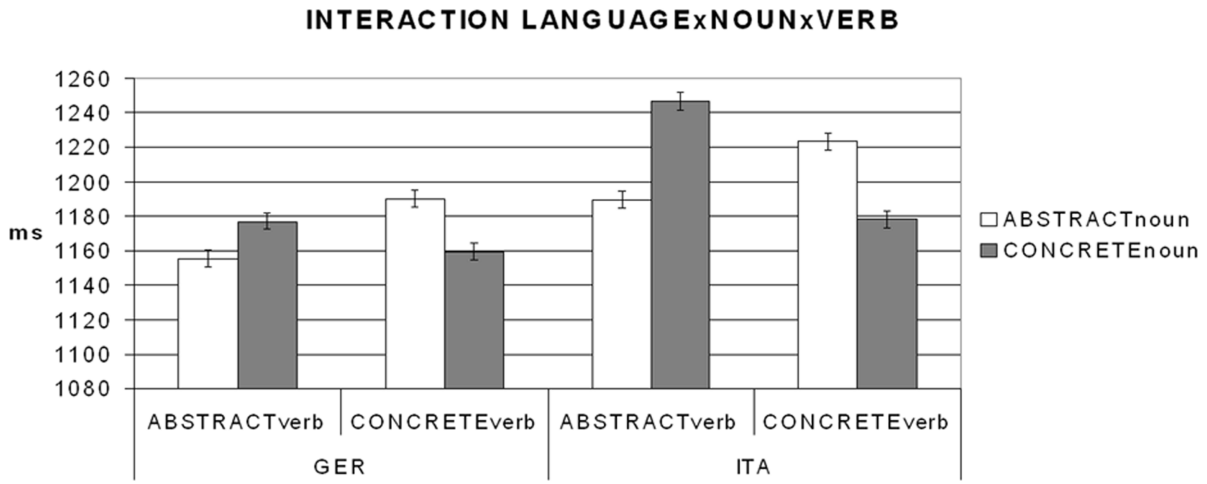


FIGURE 5 | The interaction among the kind of Language, the kind of Noun, and the kind of Verb, *p < 0.03.* *German participants were faster with Abstract Verb plus Concrete Noun, than with Concrete Verb plus Abstract Noun.* The pattern of the Italian participants was opposite.

pairs and Concrete Verb – Abstract Noun pairs ($M = 10.45$ years old; SD = 2.09; $M = 10.74$ years old; SD = 2.25). Results suggest that the different age of acquisition of sentences is explained by the noun: as shown in the literature regarding single word age of acquisition, the concrete noun is learned before the abstract one. Consistently, we found that sentences containing a concrete noun, even if in combination with an abstract verb, are acquired earlier than sentences containing an abstract noun.

RESULTS

Neither a main effect of the kind of Mapping nor a main effect of the Language used was found. Crucially, we found a significant interaction between the kind of Noun and the kind of Verb: German and Italian participants responded faster to both kinds of congruent pairs, that is both to pairs composed of an Abstract Verb plus an Abstract Noun ($M = 1172.56$ ms) and to pairs composed of a Concrete Verb plus a Concrete Noun ($M = 1168.83$ ms). Consecutively they were slower with the mixed pairs, that is, with pairs composed of an Abstract Verb plus a Concrete Noun ($M = 1211.95$ ms) and pairs composed of a Concrete Verb plus an Abstract Noun ($M = 1206.81$ ms), $F(1, 72) = 48.83$, MSe = 2328.79, $p < 0.0001$. Interestingly, Abstract Verbs combined with Abstract Nouns did not require a longer processing time than Concrete Verbs – Concrete Nouns pairs.

We also found a significant three-way interaction between Language, kind of Noun, and kind of Verb, $F(1, 72) = 5.07$, MSe = 2328.79, $p < 0.03$, see **Figure 5**. Newman–Keuls *post hoc*

analyses showed that German participants, noun (first), verb (second), were 13.25 ms faster with Abstract Verb plus Concrete Noun pairs than with Concrete Verb plus Abstract Noun pairs; on the contrary, Italian participants, noun (second), verb (first), were 23.51 ms faster with Concrete verb plus Abstract Noun pairs than with Abstract Verb plus Concrete Noun pairs; this difference reached significance only for Italian participants, $p < 0.04$. As the syntactic construction of German and Italian is different for pairs containing a transitive verb plus an object–noun, German participants, differently from Italians, were presented with the noun preceding the verb. Results with mixed pairs indicate that participants were faster when the first word was concrete rather than when it was abstract – that is when it referred to an object on which we can perform an action involving the hands (German pairs), or to an action performed with the hands (Italian pairs). This suggests that the degree of abstractness of the word plays a more important role than its grammatical class.

Moreover, the interaction between Language and kind of Verb almost reached significance as well, $F(1, 72) = 3.68$, $MSe = 3490.70$, $p < 0.06$. German participants, noun (first), verb (second), were 8.57 ms faster with pairs containing Abstract Verbs than with pairs containing Concrete Verbs. On the contrary, Italian participants, noun (second), verb (first), were 17.42 ms slower with pairs containing Abstract Verbs than with the pairs containing Concrete Verbs. Integrating these results with those obtained previously allows us to speculate that word's concreteness vs. abstractness strongly determines the time necessary to process the sentence (three-way interaction), but also that the verb has a stronger effect than the noun.

DISCUSSION

Our study showed three main new results. First we found that both the abstract verb – abstract noun combinations and the concrete verb – concrete noun combinations were processed faster than the mixed combinations. This in itself is new, particularly considering the fact that it is well known that the sentence evaluation task we used implies accessing to deep semantic representation. Our results on mixed pairs are not predicted by the No difference explanation (view 1); instead, they are predicted by views 2 and 3, and are consistent with the idea that concrete and abstract words activate parallel systems, one relying more on purely perception and action areas, the other more on sensorimotor linguistic areas. Indeed, switching between systems implies a cost in RTs, whereas remaining within the same system does not affect performance. This effect *per se* favors theories implying multiple types of representation over strictly modal and strictly amodal theories (this issue is addressed more extensively in the second section of the discussion).

The second major result we found is the three-way interaction between Language, kind of Verb, and kind of Noun. This interaction was mainly due to the fact that Germans' and Italians' results on mixed combinations were the opposite: German participants, noun (first), verb (second), were faster with abstract verb and concrete noun combinations than with concrete verb and abstract noun combinations; Italian participants, noun (second), verb (first), showed a mirror pattern. This result can be easily accounted for if we consider that the word presentation order differed across the two languages: German participants saw the noun first and then the verb, while Italians saw the same combination in a reverse order. Thus, participants were faster when the first word shown in the sentence was a concrete one, regardless of its grammatical class (verb vs. noun) and of the spoken language (German vs. Italian; for a similar result see Paivio, 1965: differently from us, in a learning and recall task he contrasted *only* abstract and concrete nouns, rather than sentences).

The third result is the marginally significant interaction we found between Language and kind of Verb. Integrating the last two findings, it seems that the abstractness vs. concreteness of the first word – that depends on the different sentences' structures – modulates sentence processing more strongly (interaction Language × Noun × Verb) than its grammatical class. Nevertheless it seems to be also an effect of the linguistic category, as verbs are more powerful than nouns in influencing subjects' responses. Fascinatingly, this result could be in keeping with the idea that the grammatical structure of a language shapes to some extent its speakers' perception of the world (Boroditsky, 2003; Gentner, 2003; Mirolli and Parisi, 2009).

Let us now consider results from RTs together, integrating them with the results obtained from the ratings of the materials. We will discuss how each theory could account for them and the problems each theory faces. We will also provide a possible neuroanatomical explanation of the results.

1. No difference view: abstract and concrete concepts have the same core representations.
   According to both (a) amodal (e.g., Fodor, 1998) and (b) strictly modal (e.g., Barsalou, 1999) theories of concepts and words, concrete, and abstract sentences are represented in the same format (amodal vs. modal). Therefore, for both amodal and modal views we should expect no difference between the four conditions, unless these differences are explained by association degree and familiarity for amodal theories, and by imageability for modal theories.

(a) According to amodal theories the results should be explained resting on the association rate between words. Therefore, the advantage of congruent over mixed sentences should be due to a higher association rate of these pairs compared to that of the mixed combinations. To check for this possibility, we calculated the familiarity and the probability of use score averages in each condition for the 120 pairs selected for the behavioral experiment. Ratings showed that, for both German and Italian participants, the advantage of congruent combinations over the mixed pairs is not explained by a supposed higher familiarity or higher probability of use of the first.

(b) According to a strictly modal theory, results regarding RT should be explained by imageability rating. An approach based more on metaphors (Lakoff, 1987) should account for the behavioral results considering the literality ratings (that indirectly give us information on the degree of metaphoricity). Actually the advantage for the Concrete Verb – Concrete Noun combination can be explained resting on its high imageability, low metaphoricity rate, and precocious

age of acquisition. But neither the modal theory nor the approach based on metaphors was verified by our results on Abstract Verb – Abstract Noun pairs, which were neither imageable nor literal (as opposed to being metaphorical) but provoked a response that was as fast as that for Concrete Verb–Concrete Noun pairs. Finally, an approach proposing that words are grounded in perceptual and especially in motor systems (Glenberg, 1997) would predict a relationship between the behavioral data and the quantity of motion scores. This was not the case, however, as the amount of movement evoked by the sentence did not explain the pattern of results with RT. Therefore, we can conclude that neither a strictly amodal nor a strictly modal theory adequately accounts for our results.

2. Non-embodied multiple representation view and
3. Embodied multiple representation view.
   Theories based on multiple types of representation – both in their non-embodied vs. embodied version – can explain the difference between congruent and mixed pairs more easily, even if resting on different reasons, that is: (I) different kinds of formats (still assuming a transduction process: Dove, 2009), or (II) a shift between different kinds of modalities, i.e., linguistic vs. a sensorimotor coding (LASS, WAT).

The interpretation that better accommodates our results assumes that abstract words are processed predominantly in the language system and concrete words are processed in the sensorimotor system to a larger extent. If processing occurs in separate systems, then the switching between concrete and abstract would imply not only conceptual costs, but also costs connected with switching between anatomical systems working in parallel. Within each system (concrete–concrete vs. abstract–abstract) the costs remain low. Some recent pieces of evidence are in line with our results. In a brain imaging study on abstract words Rüschemeyer et al. (2007) found that the processing of verbs with motor meanings (e.g., "to grasp") differed from the processing of verbs with abstract meanings (e.g., "to think"). Motor verbs produced greater signal changes than abstract verbs in several regions within the posterior premotor, primary motor (M1), and somatosensory (S1) cortices, as well as in secondary somatosensory (S2) cortex. More crucially, our interpretation is also consistent with results obtained in a brain imaging study performed using the same paradigm as the one used in the present work (Menz et al., 2011; see also Jirak et al., 2010). Using quadruples containing every possible combination for motor/non-motor verbs and for graspable/non-graspable objects, evidence showed that all motor areas were activated by language stimuli with both concrete and abstract content; but in case of concrete verb plus concrete noun processing there was a stronger engagement of areas typically involved in planning of complex and goal-directed actions (e.g., frontal operculum). In case of abstract verb plus abstract noun combinations, instead, there was a stronger engagement of the supramarginal gyrus (SMG) – typically involved in motor planning (e.g., Tunik et al., 2008) but also during phonological and articulatory words processing (e.g., Celsis et al., 1999; Pattamadilok et al., 2010) –, as well as of the MTG – that is also recruited when performing tasks critical in communication and social interaction (Mellet et al., 1998; Binder et al., 2005; Sabsevitz et al., 2005).

3. Embodied multiple representation view.

The advantage of non-mixed combinations (AA and CC) on the mixed ones (AC and CA) rules out the No difference views but can be accounted by both the Non-embodied (2) and the Embodied versions of multiple representations views (3). In order to disentangle them, the most critical result is the advantage we found when the first word was a concrete one. A Non-embodied multiple representation view (2) has difficulties in explaining this result: since the task used in the present study is a linguistic one, it should be easier to process first words which activate linguistic information, i.e., abstract words, rather than concrete ones.

LASS AND WAT

Both LASS and WAT can explain the advantage of the first concrete word. However, the explanation based on LASS would be *a posteriori*. The argument would be that, even if the task is a linguistic one, it requires deep semantic processing, and this might require more time for abstract than for concrete words. A more straightforward explanation of the longer RTs when the first word is an abstract rather than a concrete one derives from the WAT proposal. WAT assumes that both linguistic and sensorimotor processing have the same status – coherent with the advantage of the AA and CC pairs on the mixed pairs –, and it treats the issue of concepts representation as strictly related to their acquisition, stressing the different function of linguistic label for concrete vs. abstract word meanings. So the advantage of concrete words when presented first would be due to the fact that abstract words are learnt differently from concrete ones, and often with the help of a verbal explanation (see Borghi et al., 2011). It follows that for the acquisition of abstract terms the social experience due to the presence of others explaining to us specific word meanings is particularly crucial. In support of this interpretation it is worth noting that in the linguistic materials' ratings we basically found the same patterns for Imageability and Age of acquisition for both Germans and Italians: sentences containing a concrete noun (even if in combination with an abstract verb) were the easiest to imagine, and they were acquired earlier than sentences containing abstract nouns. Conversely German and Italian participants showed different patterns as far as Metaphoricity and Quantity of Motion ratings are concerned, thus they were differently influenced by the specific linguistic milieu.

In sum

The results of our behavioral study showed that participants were faster with congruent combinations, and that with mixed combinations they were faster when the first word was a concrete one, independently of the spoken language and of the word grammatical class. Results are in line with those embodied views, such as LASS and WAT, according to which both linguistic and perception

and action experience play a role in accounting for word representation. The WAT proposal is able to explain the advantage of the first concrete word better than the LASS view, ascribing it to the fact that abstract words require more time as a consequence of their peculiar acquisition modality.

Our results have a variety of implications as to how concrete and abstract words are represented in the brain, as they suggest that linguistic and perception and action information are differently distributed in accounting for concrete and abstract meanings. Consistently with recent brain imagining study (Rüschemeyer et al., 2007; Menz et al., 2011), we hypothesize that words with concrete motor content are processed to a greater extent in the perception and action systems than words with abstract content, which in turn are processed more in the linguistic areas.

ACKNOWLEDGMENTS

Thanks to Mareike Menz and Damaris Kunze for the translation of German linguistic materials; thanks to Felice Cimatti, Arthur Glenberg, Mareike Menz, and Domenico Parisi for the useful discussions. This work was supported by the European Community's Seventh Framework Programme FP7/2007-2013 – Challenge 2-Cognitive Systems, Interaction, and Robotics -, project ROSSI: Emergence of communication in RObots through Sensorimotor and Social Interaction (Grant agreement no. 216125).

REFERENCES

Aziz-Zadeh, L., Wilson, S., Rizzolatti, G., and Iacoboni, M. (2006). A comparison of premotor areas activated by action observation and action phrases. *Curr. Biol.* 16, 1818–1823.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645.

Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). "Language and simulation in conceptual processing," in *Symbols, Embodiment, and Meaning,* eds M. De Vega, A. M. Glenberg, and A. C. Graesser (Oxford: Oxford University Press), 245–284.

Barsalou, L. W., and Wiemer-Hastings, K. (2005). "Situating abstract concepts," in *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought,* eds D. Pecher and R. Zwaan (New York: Cambridge University Press), 129–163.

Binder, J. R., Westbury, C. F., McKiernam, K. A., Possing, E. T., and Medler, D.A. (2005). Distinct brain systems for processing concrete and abstract concepts. *J. Cogn. Neurosci.* 17, 905–917.

Bloom, P. (2000). *How Children Learn the Meanings of Words.* Cambridge, MA: MIT Press.

Boot, I., and Pecher, D. (2011). Representation of categories: metaphorical use of the container schema. *Exp. Psychol.* 58, 167–170.

Borghi, A. M., and Cimatti, F. (2009). "Words as tools and the problem of abstract words meanings," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Amsterdam: Cognitive Science Society), 2304–2309.

Borghi, A. M., Flumini, A., Cimatti, F., Marocco, D., and Scorolli, C. (2011). Manipulating objects and telling words: a study on concrete and abstract words acquisition. *Front. Psychol.* 2:15. doi: 10.3389/fpsyg.2011.00015

Borghi, A. M., and Pecher, D. (2011). Introduction to the special topic Embodied and Grounded Cognition. *Front. Psychol.* 2:187. doi: 10.3389/fpsyg.2011.00187

Boroditsky, L. (2003). "Linguistic relativity," in *Encyclopedia of Cognitive Science*, ed. L. Nadel (London: Macmillan), 917–922.

Brysbaert, M., Van Wijnendaele, I., and De Deyne, I. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychol. (Amst.)* 104, 215–226.

Carey, S. (1978). "The child as a word learner," in *Linguistic Theory and Psychological Reality,* eds M. Halle, J. Bresnan, and G. Millen (Cambridge, MA: MIT Press), 264–293.

Celsis, P., Boulanouar, K., Doyon, B., Ranjeva, J. P., Berry, I., Nespoulous, J. L., and Chollet, F. (1999). Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *Neuroimage* 9, 135–144.

Chatterjee, A. (2010). Disembodying cognition. *Lang. Cogn.* 2, 79–116.

Colombo, L., and Burani, C. (2002). The influence of age of acquisition, root frequency, and context availability in processing nouns and verbs. *Brain Lang.* 81, 398–411.

Desai, R. H., Binder, J. R., Conant, L. L., and Seidenberg, M. S. (2010). Activation of sensory-motor areas in sentence comprehension. *Cereb. Cortex* 20, 468–478.

Dove, G. (2009). Beyond conceptual symbols. A call for representational pluralism. *Cognition* 110, 412–431.

Dove, G. (2010). An additional heterogeneity hypothesis. *Behav. Brain Sci.* 33, 209–210.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D. Jr., Redfern, B. B., and Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition* 92, 145–177.

Fischer, M., and Zwaan, R. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol.* 61, 825–850.

Fodor, J. A. (1998). *Concepts. Where Cognitive Science went Wrong.* Oxford: Oxford University Press.

Gallese, V. (2008). Mirror neurons and the social nature of language: the neural exploitation hypothesis. *Soc. Neurosci.* 3, 317–333.

Gentner, D. (2003). "Why we're so smart," in *Language in Mind* eds D. Gentner and S. Goldin-Meadow (Cambridge, MA: MIT Press), 195–235.

Ghio, M., and Tettamanti, M. (2010). Semantic domain-specific functional integration for action-related vs. abstract concepts. *Brain Lang.* 112, 223–232.

Gibbs, R., and Steen, G. (1999). *Metaphor in Cognitive Linguistics.* Amsterdam: John Benjamins.

Gilhooly, K. J., and Gilhooly, M. L. (1980). The validity of age-of-acquisition ratings. *Br. J. Psychol.* 71, 105–110.

Glenberg, A. M. (1997). What memory is for. *Behav. Brain Sci.* 20, 1–55.

Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., and Buccino, G. (2008). Processing abstract language modulates motor system activity. *Q. J. Exp. Psychol.* 61, 905–919.

Griffiths, T. L., Steyvers, M., and Firl, A. (2007). Google and the mind predicting fluency with pagerank. *Psychol. Sci.*18, 1069–1076.

Jirak, D., Menz, M., Buccino, G., Borghi, A. M., and Binkofski, F. (2010). Grasping language. A short story on embodiment. *Conscious. Cogn.* 19, 711–720.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* Chicago: University of Chicago Press.

Lewis, M. B. (1999). Age of acquisition in face categorisation: is there an instance-based account? *Cognition* 71, B23–B39.

Louwerse, M. M., and Jeauniaux, P. (2008). "Language comprehension is both embodied and symbolic," in *Symbols and Embodiment: Debates on Meaning and Cognition,* eds M. de Vega, A. Glenberg, and A. C. Graesser (Oxford: Oxford University Press), 309–326.

Mellet, E., Tzourio, N., Denis, M., Mazoyer, B. (1998). Cortical anatomy of mental imagery of concrete nouns based on their dictionary definition. *NeuroReport* 9, 803–808.

Menz, M., Scorolli, C., Borghi, A. M., and Binkofski, F. (2011). *An fMRI Study on Abstract and Concrete Sentences Processing.* Available at: http://www.rossiproject.eu/

Mestres-Missé, A., Munte, T. F., and Rodriguez-Fornells, A. (2009). Functional neuroanatomy of contextual acquisition of concrete and abstract words. *J. Cogn. Neurosci.* 21, 2154–2171.

Mirolli, M., and Parisi, D. (2009). Towards a Vygotskyan cognitive robotics: the role of language as a cognitive tool. *New Ideas in Psychology.* doi: 10.1016/j.newideapsych.2009.07.001

Nelson, D. L., and Schreiber, T. A. (1992). Word concreteness and word structure as independent determinants of recall. *J. Mem. Lang.* 31, 237–260.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia* 9, 91–113.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). *The PageRank Citation Ranking: Bringing Order to the Web (Tech. Rep.).* Stanford, CA: Stanford Digital Library Technologies Project.

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *J. Verbal Learn. Verbal Behav.* 4, 32–38.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. New York: Oxford University.

Pattamadilok, C., Knierim, I. N., Kawabata Duncan, K. J., and Devlin, J. T. (2010). How does learning to read affect speech perception? *J. Neurosci.* 30, 8435–8444.

Pecher, D., Boot, I., and van Dantzig, S. (2011). "Abstract concepts: sensory-motor grounding, metaphors, and beyond," in *The Psychology of Learning and Motivation,* Vol. 54, ed. B. Ross (Burlington: Academic Press), 217–248.

Pecher, D., Zeelenberg, R., and Barsalou, L. W. (2003). Verifying properties from different modalities for concepts produces switching costs. *Psychol. Sci.* 14, 119–124.

Pezzulo, G., and Castelfranchi, C. (2007). The symbol detachment problem. *Cogn. Process.* 8, 115–131.

Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis.* Cambridge, MA: MIT Press.

Pulvermüller, F. (in press). Meaning and the brain: the neurosemantics of referential, interactive, and combinatorial knowledge. *J. Neurolinguist.* doi: 10.1016/j.jneuroling.2011.03.004

Rüschemeyer, S. A., Brass, M., and Friederici, A. D. (2007). Comprehending prehending: neural correlates of processing verbs with motor stems source. *J. Cogn. Neurosci.* 19, 855–865.

Sabsevitz, D. S., Medler, D. A., Seidenberg, M., and Binder, J. R. (2005). Modulation of the semantic system by word imageability. *Neuroimage* 27, 188–200.

Schwanenflugel, P. (1991). "Why are abstract concepts hard to understand?" in *The Psychology of Word Meanings,* ed. P. Schwanenflugel (Hillsdale, NJ: Erlbaum), 223–250.

Scorolli, C., Jacquet, P.O., Binkofski, F., Nicoletti, R., Tessari, A., and Borghi, A. (2011). "Involvement of primary motor cortex in abstract and concrete sentences processing," in *Embodied and Situated Language Processing*, Bielefeld.

Sha, G. (2010). Using Google as a super corpus to drive written language learning: a comparison with the British National Corpus. *Comput. Assist. Lang. Learn.* 23, 377–393.

Tunik, E., Lo, O. Y., and Adamovich, S. V. (2008). Transcranial magnetic stimulation to the frontal operculum and supramarginal gyrus disrupts planning of outcome-based hand–object interactions. *J. Neurosci.* 28, 14422–14427.

Turken, A. U., and Dronkers, N. F. (2011). The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Front. Syst. Neurosci.* 5:1. doi: 10.3389/fnsys.2011.00001

Wauters, L. N., Tellings, A. E. J. M., Van Bon, W. H. J., and Van Haaften, A. W. (2003). Mode of acquisition of word meanings: the viability of a theoretical construct. *Appl. Psycholinguist.* 24, 385–406.

Wiemer-Hastings, K., Krug, J., and Xu, X. (2001). "Imagery, context availability, contextual constraints and abstractness," in *Proceedings of 23rd Annual Meeting of the Cognitive Science Society*, eds J. D. Moore and K. Stenning (Hillsdale, NJ: Lawrence Erlbaum Associates), 1106–1111.

Zevin, J. D., and Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *J. Mem. Lang.* 47, 1–29.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 March 2011; paper pending published: 06 July 2011; accepted: 25 August 2011; published online: 15 September 2011.

Citation: Scorolli C, Binkofski F, Buccino G, Nicoletti R, Riggio L and Borghi AM (2011) Abstract and concrete sentences, embodiment, and languages. Front. Psychology ***2****:227. doi: 10.3389/fpsyg.2011.00227*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 18 January 2012
doi: 10.3389/fpsyg.2011.00384

From reference to sense: how the brain encodes meaning for speaking

Laura Menenti [1,2] ***, Karl Magnus Petersson*** [1,3] ***and Peter Hagoort*** [1,3] ***

[1] *Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands*
[2] *Institute for Neuroscience and Psychology, University of Glasgow, Glasgow, UK*
[3] *Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Ken McRae, University of Western Ontario, Canada
Mante Nieuwland, Basque Center on Cognition, Brain and Language, Spain

***Correspondence:**
Laura Menenti, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, Netherlands.
e-mail: laura.menenti@mpi.nl;
Peter Hagoort, Donders Institute for Brain, Cognition and Behaviour, Donders Centre for Cognitive Neuroimaging, Radboud University Nijmegen, P.O. Box 9101, 6500 HB Nijmegen, Netherlands.
e-mail: peter.hagoort@donders.ru.nl

In speaking, semantic encoding is the conversion of a non-verbal mental representation (the reference) into a semantic structure suitable for expression (the sense). In this fMRI study on sentence production we investigate how the speaking brain accomplishes this transition from non-verbal to verbal representations. In an overt picture description task, we manipulated repetition of sense (the semantic structure of the sentence) and reference (the described situation) separately. By investigating brain areas showing response adaptation to repetition of each of these sentence properties, we disentangle the neuronal infrastructure for these two components of semantic encoding. We also performed a control experiment with the same stimuli and design but without any linguistic task to identify areas involved in perception of the stimuli *per se*. The bilateral inferior parietal lobes were selectively sensitive to repetition of reference, while left inferior frontal gyrus showed selective suppression to repetition of sense. Strikingly, a widespread network of areas associated with language processing (left middle frontal gyrus, bilateral superior parietal lobes and bilateral posterior temporal gyri) all showed repetition suppression to both sense and reference processing. These areas are probably involved in mapping reference onto sense, the crucial step in semantic encoding. These results enable us to track the transition from non-verbal to verbal representations in our brains.

Keywords: semantics, conceptual representation, language production, fMRI, fMRI adaptation

INTRODUCTION

Look at that guy hitting the other guy! After reading this sentence, you presumably have a mental representation of two adult male persons, of whom one is hitting the other. They are both male and adult but they are still two different persons. A linguistic distinction within the domain of semantics, is the difference between *reference* and *sense* of a linguistic expression (Frege, 1892). The sense of an expression is its linguistic meaning, the reference is the entity the expression refers to. In the representation of the sentence *Look at that guy hitting the other guy!* there are two *guys* (for instance a blond and a dark-haired guy, as indicated by "that" and "the other"), but they are both referred to by the same sense, the word g*uy* (an adult male person). This sense thus has two possible references. The reverse is also possible. If you knew more about the two *guys* you might be shouting: *Look at that man hitting his son!* in the same situation. *His son* and *the other guy* are then two possible senses which can have the same referent.

In the view of Jackendoff (2002), which we adopt in the current paper, referents are representations in our minds. For concrete objects they are representations in our minds, of objects in the real world, constructed by the perceptual system. These representations are considered concepts, which thus are non-linguistic in nature. Sense, then, is that part of meaning that is encoded in the form of the utterance. In other words, sense (linguistic meaning) is the interface between the conceptual system and linguistic form (spanning both phonology and syntax; Jackendoff, 2002). Speaking is the conversion of an intention to communicate a message into a linearized string of speech sounds. An essential step in this process is semantic encoding – the retrieval of the relevant concepts and the specification of semantic structure (Levelt, 1989). In this step, the intended reference needs to be mapped onto a sense, for it to be expressed. In this mapping process, certain semantic choices have to be made, such as referring to the entities in the referential domain by, for instance, "the guy" or "the man on the chair." From a processing point of view, then, reference forms the input to semantic encoding, while sense is the output. Semantic encoding itself is the computation necessary to map reference (the input) onto the sense (the output) in order to generate the appropriate output. In this paper, we consider sense to be equivalent to the *preverbal message* in sentence production (Levelt, 1989). The preverbal message is the semantic structure that forms the output of semantic encoding and the input to syntactic and phonological encoding.

In speaking establishing reference is the first step of semantic encoding, necessary to utter a sentence in the first place. As few neuroimaging studies investigating semantic encoding in sentence production have so far been conducted, in this fMRI study we aim to fill that gap. Picture naming paradigms have previously been used in fMRI albeit in single word studies. Retrieving a name for a picture has been shown to involve more activity in bilateral temporal areas, the left frontal lobe, bilateral occipital areas, bilateral parietal areas, and the anterior cingulate

(Kan and Thompson-Schill, 2004) than does making visual decisions about abstract pictures. A similar set of areas has been shown to increase activity in picture naming and reading aloud compared to counting (Parker Jones et al., 2011). These data suggest that a large network of areas may be involved in semantic encoding. Also, while part of this network are well-established language areas, some are not. Perhaps, then, these are areas encoding the reference for these materials.

Moving on to sentence production, in a previous fMRI adaptation study on sentence-level processing, we compared the neuronal structure underlying computation of semantic structure of an utterance in comprehension and production (Menenti et al., 2011). More specifically, we investigated the construction of *thematic role structure*, the relation between the different concepts and events, or "who does what to whom." This aspect of semantic structure forms a crucial interface between conceptual structure (the domain of reference) and syntactic structure (the grammatical roles). Schematically a thematic role structure can be stated as a predicate with arguments: ROB(THIEF, LADY(OLD)). There is a "ROB" event, performed by a THIEF (the agent of the action) to the expenses of a LADY (the patient of the action), who has the property of being OLD. In our study, photographs depicting transitive events (events requiring an agent and a patient, such as ROB, KISS, HIT) provided the context for the sentences, which were either produced or heard by the participants. We found bilateral posterior middle temporal gyri involved in this component of sentence processing.

While this study provided valuable insights on the neuronal infrastructure underlying different steps in sentence production and comprehension, the semantic encoding manipulation disregarded the distinction between reference and sense. The next question, and the one underlying the present study, then, is how the different areas involved in semantic encoding play a role in processing the input (reference), the output (sense), and the process of mapping the one onto the other. In this sentence production fMRI adaptation study we again focused on thematic role structure as an essential part of semantic structure. In a picture description paradigm, we manipulated repetition of semantic structure across subsequent sentences, crossing repetition of reference and sense.

Our paradigm involved pictures of transitive events being enacted by two actors. We operationalized sense as the literal sentence used to describe the picture. Reference we considered the sum of the action involved, the roles of actors as agents and patients, and the exact spatial configuration of agents and patients.

In our task, the actors in the picture were colored and these colors varied for the same depicted situation. Participants could therefore subsequently describe the same situation as "The yellow man hits the blue woman." and then as "The red man hits the green woman." Although the picture therefore looked slightly different in the two trials, the colors were an arbitrarily varying property of the individuals in the picture and the participants were made aware of this (see Materials and Methods). We do not consider such arbitrary variations to be part of reference. One might consider this parallel to the fact that we change clothes every day: they make us look different but do not thereby cause us to become different individuals. The reference of the expression was therefore kept constant but the sense changed. In a complementary fashion, the sentence "The red man hits the green woman." could be used in subsequent trials to describe a different hitting event involving different participants. Sense was kept constant, but the reference changed. This allowed us to distinguish the situation the participants spoke about from the utterance they used to speak about it.

As can be seen in **Figure 1**, this means that our relevant "novel reference" condition still has considerable overlap with the prime. We chose this approach to eliminate any potential confounds. For instance, repeating the actors between prime and "repeated reference" target but not between prime and "novel reference" target would leave effects open to, for instance, the alternative interpretation that we are looking at face repetition effects. By choosing the most narrow comparison possible, we can be more sure of the interpretation of the results, while admittedly running the risk of missing some other potentially relevant effects.

To further investigate the distinction between non-verbal and verbal processing of meaning, we performed a control experiment. In this experiment we showed participants the exact same stimulus sequences, but this time paired with a non-linguistic task. Any brain areas involved in processing only the non-linguistic, conceptual representations involved in interpreting the pictures (i.e., the reference), should also show an adaptation effect without a linguistic task. On the other hand, brain areas involved in converting meaning into language (the sense), should not show adaptation effects in such a setting.

Our hypothesis was that areas involved in processing the conceptual input to semantic encoding should show adaptation effects for repetition of reference in both the speaking and control experiments, while not showing sensitivity to repetition of sense. Areas involved in semantic encoding itself, that is, in mapping reference onto sense, should show adaptation to repetition of both reference and sense. Finally, areas involved in processing the output of semantic encoding, the sense, should show adaptation to repetition of sense in the speaking experiment, and should not show sensitivity to repetition of reference.

MATERIALS AND METHODS

PARTICIPANTS

Twenty-four right-handed subjects took part in each experiment (speaking: 12 female; mean age 25.2 years, SD 7.5; control: 13 female; mean age 22.8 years, SD 3.17). All subjects were healthy native Dutch speakers with normal or corrected-to-normal vision and had attended or were attending university education in the Netherlands. All subjects gave written informed consent prior to the experiment and received a fee or course credit for their participation. No participants took part in both experiments.

STIMULI

Our target stimulus set contained 1152 photographs that depicted 36 transitive events such as *kiss, help, strangle* with the agent and patient of this action. Four couples performed each action (2 × men/women; 2 × boy/girl), in two configurations (one with the man/boy as the agent and one with the woman/girl). These

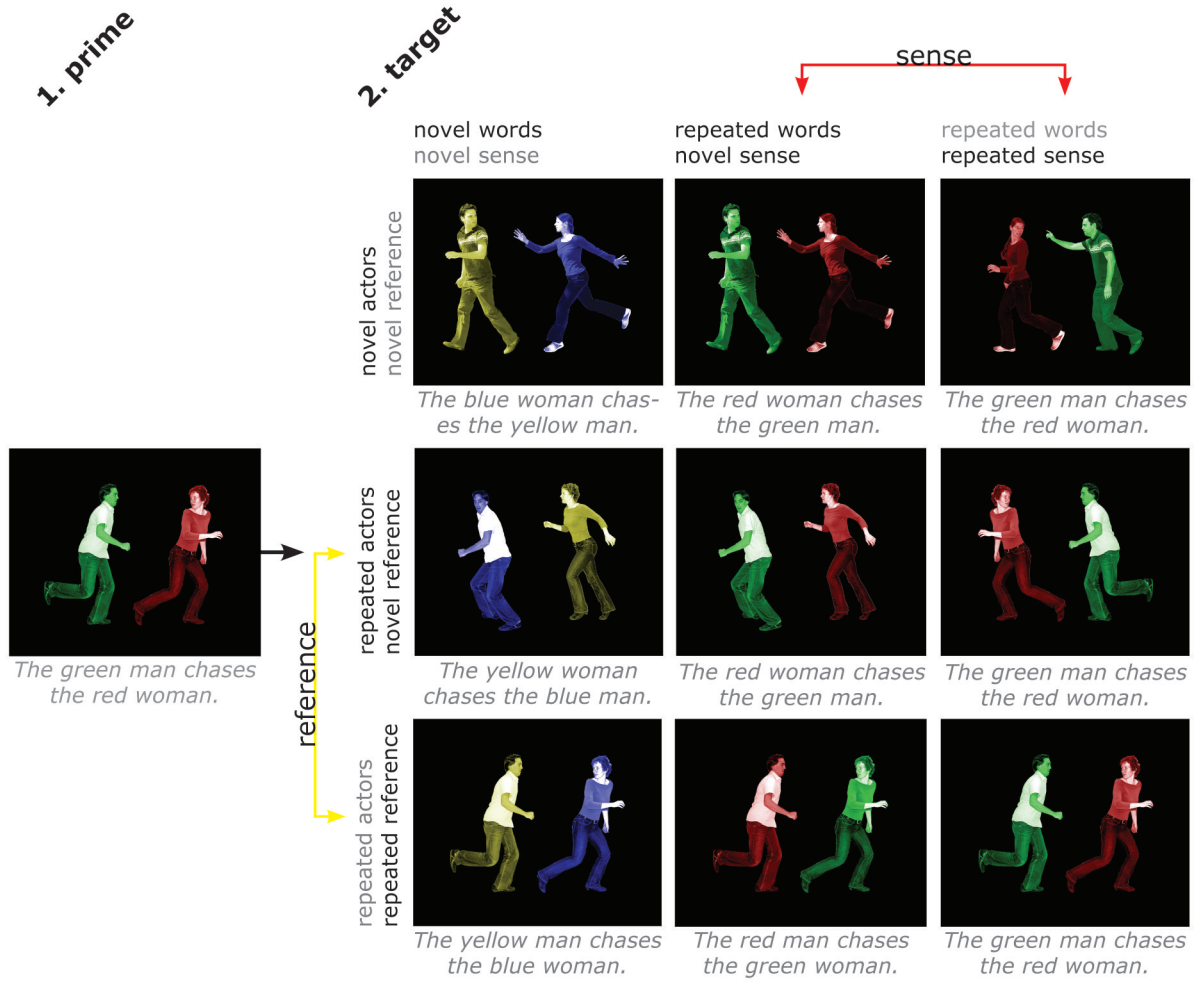


FIGURE 1 | Design and stimuli. Subjects described scenes with short sentences containing action, agent, and patient and their respective colors. There are four factors (words, sense, actors, reference), with two levels (repeated/novel) each. Separate repetition of sense and individual words allows distinguishing areas that are sensitive to the overall meaning of a sentence from those that are sensitive to repetition of words but not to repetition of sentence meaning. Separate repetition of actors and reference allows distinguishing areas that are sensitive to the overall situation from those that are sensitive to the people involved irrespective of the specific situation. The four factors cannot be fully crossed, since it is not possible to repeat sense without, at least partly, repeating words, as it is not possible to repeat a situation while changing the people involved. Therefore, each of the factors varies only at one level of the factor it could potentially be confounded with. The contrasts of interest are between the middle and right column for sense (hence keeping repetition of words controlled) and between the middle and bottom row for reference (hence keeping repetition of actors controlled). Likewise, the effect of word repetition was computed by comparing the leftmost and middle column, and the effect of actor repetition by comparing the top and middle row.

$36 \times 4 \times 2$ pictures were further edited so that the agent and patient each had a different color (red–green, green–red, blue–yellow, yellow–blue), and these $36 \times 4 \times 2 \times 4$ pictures were also flipped so that the position of the agent could be either left or right on the picture. The filler stimuli contained pictures depicting 868 intransitive (e.g., *The boy runs.*) and 160 locative (e.g., *The balls lie on the table.*) events. The actors and objects in these pictures were also colored in red, green, yellow or blue. The control experiment further included catch stimuli, which constituted 10% of the trials. These were pictures similar to the target pictures, but containing a range of visual defects that the subjects had to detect. The stimuli are available for use from the authors.

DESIGN

The design is illustrated in **Figure 1**, and was identical for both experiments. There were four factors (words, sense, actors, reference), with two levels (repeated/novel) each. Contrasting repetition of sense and individual words allowed us to distinguish areas that are sensitive to the overall meaning of a sentence, and those that are sensitive to repetition of words but not to repetition of sentence meaning. The verb and nouns were always repeated for target trials, and only the adjectives could vary. This was necessary due to the constraints on repetition of elements in the pictures for the different conditions. For instance, since "repeated reference" entailed repeating both the action and the people involved, this meant also repeating the words used to refer to these elements.

Contrasting repetition of actors and reference allowed us to distinguish areas that are sensitive to the overall situation from those that are sensitive to the people involved irrespective of the specific situation. The four factors could not be fully crossed, since it is not possible to repeat sense without, at least partly, repeating words, like it is not possible to repeat an event (the reference) while changing the people involved. Therefore, we performed the relevant comparison for each factor at only one level of the factor it could potentially be confounded with (see **Figure 1**).

The target items were presented in 78 mini-blocks with an average length of 5.4 items (range 3–7 items). The target blocks were alternated with filler blocks, with an average length of 3.5 items. Filler blocks served the purpose of increasing variability in syntactic structures, words, and visual properties of the sentences and pictures. Subjects were unaware of the division in blocks, as the items were presented at a constant rate. We used a running priming paradigm where each target item also served as prime for the subsequent target item. No condition was repeated twice in a row. Since there were 78 target blocks, 78 transitive items (the first of each mini-block) served as primes only. The remaining 315 transitive items (2–6 per block depending on block length) constituted the target trials so that there were 35 items per condition. Each subject saw a different randomized list, which consisted of 393 transitive (78 prime-only and 315 target items) stimuli and 262 filler stimuli. For the speaking experiment, these were randomly sampled from the 868 intransitive and 160 locative pictures in the filler stimulus set. In the control experiment, the 262 pictures were always 65 catch (10% of total number of trials), 67 locative and 130 intransitive pictures.

TASK AND PROCEDURE

Speaking experiment: participants first read the instructions and were given the opportunity to ask questions. The instructions not only explained the task, but also introduced all the different frequently occurring actors as separate individuals, along with the same photo of them in every color. This way, we made sure that the participants were aware that the colors were arbitrarily varying properties of the different actors.

Each target picture was preceded by its corresponding verb. Participants described the picture with a short sentence, using the presented verb. In this sentence they had to mention both persons and their colors. The experiment consisted of two runs of 39 min. This served the purpose of not keeping participants in the MRI-scanner for too long; the runs were otherwise completely equivalent. The participants underwent a 5-min anatomical scan after the first run, and were then taken out of the MR-scanner for a break before they underwent the second run. The responses were recorded in order to extract reaction times (RTs). The experimenter coded the participant's responses online for correctness and prevoicing. Prevoicing was coded to ensure correct measurement of RTs, which were extracted through thresholding of the speech recording (see below for details). Each trial lasted 7000 ms and consisted of the following events: the verb was presented with a jittered start time of 0–1000 ms after the start of the trial, and a duration of 500 ms. After an ISI of 500–2500 ms the picture was presented for 2000 ms.

Control experiment: participants first read the instructions and were given the opportunity to ask questions. The participant's task was to act as a "proof viewer" scanning a set of pictures for misprints. They were given examples of both correct pictures and possible misprints. They were instructed to press a button whenever they detected a misprint, and to do nothing if the pictures were ok. The experiment consisted of two runs of 22 min. The participants underwent a 5-min anatomical scan between runs. Each trial lasted 4000 ms, in which the picture was displayed with a jittered start time of 0–1500 ms from trial onset, and stayed on screen for 1000 ms. We chose different timing parameters for this experiment, to avoid it becoming incredibly boring.

DATA ACQUISITION AND ANALYSIS

Data acquisition took place in a 3-T Siemens Magnetom Tim-Trio MRI-scanner. Participants were scanned using a 12-channel surface coil. To acquire our functional data we used parallel-acquired inhomogeneity-desensitized fMRI (Poser et al., 2006). This is a multi-echo EPI: images are acquired at multiple TE's following a single excitation. The TR was 2398 ms and each volume consisted of 31 slices of 3 mm thickness with a slice-gap of 17%. The voxel size was 3.5 mm × 3.5 mm × 3 mm and the field of view was 224 mm. Functional scans were acquired at multiple TEs following a single excitation (TE_1 at 9.4 ms, TE_2 at 21.2 ms, TE_3 at 33 ms, TE_4 at 45 ms, and TE_5 at 56 ms with echo spacing of 0.5 ms) so that there was a broadened T_2^* coverage. Because T_2^* mixes into the five echoes in a different way, the estimate of T_2^* is improved. Accelerated parallel imaging reduces image artifacts and thus is a good method to acquire data when participants are producing sentences in the scanner (causing motion and susceptibility artifacts). The number of slices did not allow acquisition of a full brain volume in most participants. We always made sure that the entire temporal and frontal lobes were scanned because these were the areas where the fMRI adaptation effects of interest were expected. This meant however that data from the superior posterior frontal lobe and the anterior superior parietal lobe (thus data from the top of the head) were not acquired in several participants. The functional scans of the first and second runs were aligned using AutoAlign. A whole-brain high resolution structural T1-weighted MPRAGE sequence was performed to characterize participants' anatomy (TR = 2300 ms, TE = 3.03 ms, 192 slices with voxel size of 1 mm^3, FOV = 256), accelerated with GRAPPA parallel imaging (Griswold et al., 2002).

For the behavioral data of the speaking experiment, to separate participants' speech from the scanner sound and extract RTs, the speech recordings were bandpass filtered with a frequency band of 250–4000 Hz and smoothed with a width half the sampling rate. Response onsets and durations were determined through thresholding of these filtered recordings (basically, a *post hoc* voicekey) and linked to the stimulus presentation times to extract the RTs and total speaking times. Trials with errors or prevoicing were discarded from the analysis. The planning times, speaking times and total response times for correct responses to the target items were analyzed in a repeated measures ANOVA using SPSS.

The fMRI data were preprocessed and analyzed with SPM5 (Friston et al., 1995). The first 5 images were discarded to allow for T_1 equilibration. Then the five echoes of the remaining images

were realigned to correct for motion artifacts (estimation of the realignment parameters was done for one echo and then copied to the other echoes). Subsequently the five echoes were combined into one image with a method designed to filter task-correlated motion out of the signal (Buur et al., 2009). First, echo two to five (i.e., TE_2, TE_3, TE_4, and TE_5) were combined using a weighting vector dependent on the measured differential contrast to noise ratio per voxel. The time course of an image acquired at a very short echo time (i.e., TE_1) was used as a voxelwise regressor in a linear regression for the combined image of TE_2, TE_3, TE_4, and TE_5. Weighting of echoes was calculated based on 25 volumes acquired before the actual experiment started. The resulting images were coregistered to the participants' anatomical scan, normalized to MNI space and spatially smoothed using a 3D isotropic Gaussian smoothing kernel (FWHM = 8 mm).

We then performed first- and second-level statistics. For the first level general linear model (GLM), we modeled the individual start time of the picture. The events of our model were convolved with the canonical hemodynamic response function included in SPM5. In the speaking experiment, the first level model included verbs, filler pictures, prime pictures, the nine conditions and errors. Erroneous responses were therefore put in a separate regressor, leaving only correct responses in the actual analyses. For the control experiment, the first level model included filler pictures, prime pictures, the nine conditions, and catch trials. Both models included the six motion parameters as event-related regressors of no interest. The second-level model consisted of a 9 (condition) × 2 (experiment) factorial design. All effects were then tested by computing the appropriate contrasts for the model. We performed two types of analyses to test our hypotheses: to find intersections between different effects, we conducted conjunction analyses. In these analyses multiple different contrasts are tested, and only areas showing an effect in all tested contrasts under a conjunction null hypothesis result in a significant conjunction (Friston et al., 2005). To look for areas sensitive to one factor but not the other, we applied exclusive masking. In such an analysis, the significant clusters for one factor are overlaid with a low-threshold mask for the other factor ($p < 0.20$ uncorrected voxelwise), and only clusters that survive the masking procedure are reported. Note that due to the very nature of the type of statistical framework we employ, we cannot prove that an effect does *not* exist. However, if an effect does not survive thresholding at $p < 0.20$ uncorrected voxelwise, it may be said to be very weak at the very least. For all tests, the cluster size at voxelwise threshold $p < 0.001$ uncorrected was used as the test statistic and only clusters significant at $p < 0.05$ corrected for multiple non-independent comparisons are reported. Local maxima are also reported for all clusters with their respective voxelwise family wise error (FWE) corrected p-values. The effects for repetition of words and actors are reported in the tables, but since the aim of the study is to distinguish reference and sense we focus on those two factors in discussing the results.

RESULTS

BEHAVIORAL DATA

For the speaking experiment, we performed repeated measures GLMs on the planning times (RTs), speaking times (the duration of the response), and the total planning + speaking times. The model included one factor (condition) with 9 levels, and the three dependent variables. The effects reported were computed through custom hypothesis tests within this model, using contrasts much like for the fMRI analyses. The data are reported in **Figure 2**. For planning times, repetition of sense, actors and reference produced significant priming effects [words: $F < 1$; sense: $F(1,23) = 109.53$, $p < 0.001$; actors: $F(1,23) = 22.95$, $p < 0.001$; reference: $F(1,23) = 94.60$, $p < 0.001$]. For speaking times, repetition of reference and sense significantly affected the duration of the response [words: $F(1,23) = 1.52$, $p < 0.232$; sense: $F(1,23) = 12.31$ $p < 0.002$; actors: $F(1,23) = 3.71$, $p < 0.066$; reference: $F(1,23) = 9.50$, $p < 0.005$]. However, the direction of these effects was reversed. Priming led to shorter planning times but longer speaking times. Analyses on the total time the participants took to complete the response (so planning plus speaking time) again revealed significant effects for reference and sense [words: $F(1,23) < 1$; sense: $F(1,23) = 13.41$, $p < 0.001$; actors: $F(1,23) = 2.78$, $p < 0.11$; reference: $F(1,23) = 33.307$, $p < 0.001$]. The total response time mirrored the planning time pattern: when primed, subjects were faster to complete the response. There were no significant interactions in any of the analyses, in so far as these could be computed given the design. In the control experiment, the average d-prime was 0.7, indicating that participants did pay attention.

fMRI RESULTS

All results are reported in **Tables 1** and **2**, and depicted in **Figure 3**. **Table 2** lists the main effects for all factors in the design; we limit the discussion to the more specific results for reference and sense as listed in **Table 1**.

Speaking experiment

The first step in semantic encoding, the input, is to compute a non-linguistic representation underlying the sentence to be produced. We therefore looked for areas exhibiting fMRI adaptation to reference, while not showing an effect of sense repetition. The BOLD-response of the bilateral temporo-parietal junctions (BA 39/37/19) and the precuneus decreased after repetition of reference. In the right middle frontal and inferior gyrus (BA 45/46) the BOLD-response increased after repetition of reference.

The next step in semantic encoding, is to map the reference onto a linguistic semantic structure that can be expressed, the sense. This is the actual process of semantic encoding. We therefore tested for areas sensitive to repetition of both reference and sense. The bilateral superior parietal lobes (BA 7), fusiform gyrus (BA 37) and posterior middle temporal areas showed suppression in the conjunction analysis. The left calcarine sulcus (BA 17) exhibited suppression as well. Finally, three frontal clusters in the left middle frontal gyrus, left SMA and left precentral gyrus (all BA 6) also showed repetition suppression. We also tested for increased responses upon repetition (enhancement), but found no areas exhibiting this pattern.

Finally, the mapping process produces an output, the sense. This should be reflected in regions showing fMRI adaptation to sense, without showing an effect for reference. The BOLD-response of the pars triangularis in the left inferior frontal gyrus (BA 45) was reduced after repetition of sense. The response in the left angular

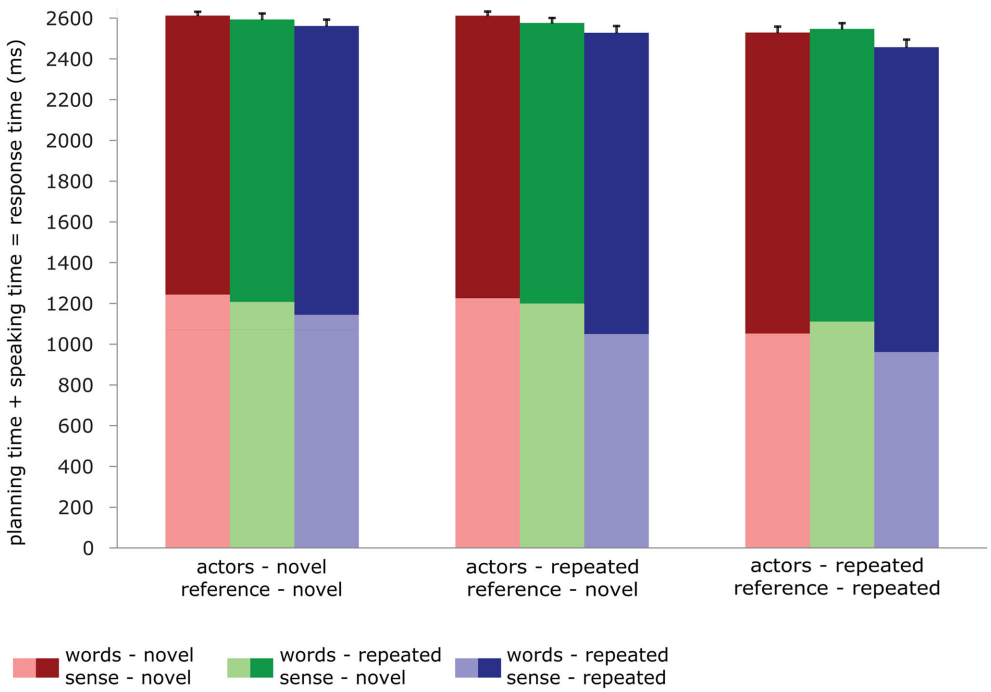


FIGURE 2 | Behavioral data in the speaking experiment: reaction times (light shades), durations (dark shades), and total speaking times (total bar length) for all conditions. Error bars represent SE of the mean of the total speaking times.

gyrus (BA 39/19) and in the left middle frontal gyrus (BA 44/9), on the other hand, increased after repetition of sense.

Control experiment

In the process of semantic encoding, the input is the conceptual representation that has to be transformed into a preverbal message. To some extent at least, such a representation should also be constructed when we are not speaking. The only significant effect in the control experiment was indeed a main effect of repetition of reference in bilateral posterior middle temporal gyri/inferior parietal lobe (BA 37/39). This effect survived masking with sense in the control experiment, and was also the same as the main effect of reference in the speaking experiment, as demonstrated by a conjunction analysis (**Table 3**). The right middle temporal gyrus (BA 21) showed enhancement upon repetition of reference.

DISCUSSION

In this sentence production study, we aimed to distinguish brain areas sensitive to reference (the mental representation an utterance refers to) and the sense (the linguistic structure that interfaces meaning with linguistic form). The behavioral data in the speaking experiment showed that both reference and sense priming affect the responses, and that these two effects do not interact. This shows that both processes are psychologically real and distinct, and that priming them affects the speed with which a sentence can be produced.

In speaking, constructing an utterance is an incremental process, involving several steps (Levelt, 1989). The first is to construct a *preverbal message*. In the present experiment, this requires encoding the situation we want to talk about (MAN_a hitting $WOMAN_b$) into a thematic role structure which can be described as HIT(MAN(YELLOW), WOMAN(BLUE)): there is a HIT event, performed by a MAN, who has the property of being YELLOW, at the expenses of a WOMAN who has the property of being BLUE (perhaps reasonable given that she is being HIT). As outlined in the introduction, the input is the reference, the output the sense. We wanted to find out which areas in the brain are involved in this mapping process. In the following, we will trace step by step how, based on our results, we think the brain comes to encode an utterance.

The first step is to build a representation of a situation we are going to talk about – the reference. This representation forms the input to semantic encoding, and is non-linguistic (conceptual) in nature. As outlined in the introduction, in the case of a concrete referent this representation is the result of perceptual processes within the perceptual system – in the present case, the visual system. Presumably, such a representation is, at least to some extent, built for what we perceive independently of whether we are going to talk about it or not. In the present paradigm, this step should be independent of the sense of the final utterance. Areas showing suppression to repetition of reference but not sense were the bilateral occipito-temporo-parietal junctions (BA 37/39/19) and the precuneus. Data from the control experiment corroborate the idea that the role of these areas in reference in the present experiment is primarily to build a perceptual representation: the same bilateral areas at the junction of the occipital, temporal and parietal lobes show suppression to repetition of reference in the absence

Table 1 | Overlap and segregation of reference and sense.

Effect	Cluster	BA	Anatomical label	Global and local maxima			Cluster-level		Voxel-level	
				x	*y*	*z*	*K*	*p*	*T*	*p* (FWE)
Sense and reference	1	37	Fusiform_L	−42	−60	−12	4162	0.000	5.95	0.000
		7	Parietal_Sup_L	−24	−62	54			5.57	0.001
		7	Parietal_Sup_L	−22	−72	44			5.33	0.004
	2	6	Frontal_Mid_L	−26	−8	50	728	0.000	5.73	0.001
		6	Precentral_L	−40	0	50			4.99	0.018
	3	19	Occipital_Inf_R	44	−76	−2	1323	0.000	4.90	0.026
		19	Occipital_Inf_R	38	−72	−8			4.86	0.030
		37	Temporal_Mid_R	46	−66	10			4.41	0.161
	4	6	Supp_Motor_Area_L	−4	10	56	191	0.035	4.76	0.045
	5	7	Parietal_Sup_R	26	−58	54	461	0.000	4.74	0.048
		7	Parietal_Sup_R	24	−72	50			4.27	0.247
		7	n/a	24	−48	48			3.94	0.583
	6	6/44	Precentral_L	−44	6	20	299	0.005	4.42	0.153
		6	Precentral_L	−42	−4	32			4.27	0.253
	7	17	Calcarine_L	−8	−92	6	329	0.003	4.08	0.423
		17	Calcarine_L	−10	−82	8			3.77	0.772
		17	Calcarine_L	−14	−68	8			3.74	0.807
Sense-suppression	1	n/a	Frontal_Inf_Tri_L	−34	18	24	186	0.039	4.71	0.054
(masked for reference)		45	Frontal_Inf_Tri_L	−50	28	26			4.46	0.136
		45	Frontal_Inf_Tri_L	−38	26	26			4.10	0.409
Sense-enhancement	1	39	Angular_L	−54	−58	36	497	0.000	5.85	0.000
(masked for reference)		19	Occipital_Mid_L	−42	−74	38			4.23	0.283
	2	9	Frontal_Mid_L	−26	24	44	251	0.012	4.03	0.476
		44	Frontal_Mid_L	−42	18	40			4.01	0.501
Reference-suppression	1	39	Temporal_Mid_R	50	−66	20	805	0.000	9.44	0.000
(masked for sense)		39	Occipital_Mid_R	42	−74	26			8.09	0.000
		37	Temporal_Mid_R	60	−60	12			7.81	0.000
	2	39	Temporal_Mid_L	−42	−68	20	617	0.000	8.99	0.000
		39	Temporal_Mid_L	−54	−66	18			7.37	0.000
		19	Occipital_Mid_L	−30	−78	32			5.35	0.002
	3	23	Precuneus_R	4	−58	24	1389	0.000	5.56	0.001
		n/a	Precuneus_R	2	−54	40			5.24	0.006
		n/a	Precuneus_L	−10	−50	52			5.16	0.009
Reference-enhancement	1	45	Frontal_Mid_R	38	46	10	416	0.001	4.79	0.040
(masked for sense)		46	Frontal_Inf_Orb_R	44	48	−4			4.31	0.218

Listed are the MNI-coordinates for the first three local maxima for each significant cluster in the relevant comparisons ($p < 0.05$ corrected cluster-level, threshold $p < 0.001$ uncorrected voxelwise, exclusive masks $p < 0.20$ uncorrected voxel-wise). Anatomical labels are derived from the Automatic Anatomical Labeling atlas (Tzourio-Mazoyer et al., 2002) and from Brodmann's atlas. Cluster-level statistics are listed for each cluster, voxel-level statistics also for local maxima.

of a linguistic task. The finding that these areas are involved in generating the non-linguistic representation to refer to now also allows us to further specify a previous finding on semantic encoding in sentence production: in a previous study, we found that part of the superior right MTG is sensitive to sentence- but not word-level meaning (Menenti et al., submitted). This effect overlaps with the area sensitive to repeated reference but not sense, and therefore was presumably due to the encoding of the referent as well. These same regions have also been found sensitive to subsequent memory for short stories (Hasson et al., 2007), a further indication that they are involved in constructing a representation of what linguistic material refers to. Repetition of reference did not just elicit suppression: in right inferior frontal gyrus the response increased upon repeated presentation. The repetition enhancement effect for reference in right inferior frontal cortex was particularly striking since large parts of contralateral left inferior frontal cortex showed repetition suppression for reference. Repetition enhancement has been postulated to be caused, among other things, by novel network formation due to the construction of new representations (Henson et al., 2000; Conrad et al., 2007; Gagnepain et al., 2008; Segaert et al., submitted). In speech comprehension, right inferior frontal cortex has previously been implicated in the construction of a situation model (Menenti et al., 2009; Tesink et al., 2009), a mental representation of text containing

Table 2 | Main effects for all factors in the design.

Effect	Cluster	BA	Anatomical label	Global and local maxima			Cluster-level		Voxel-level	
				x	*y*	*z*	*K*	*p*	*T*	*p* (FWE)
Main effect words	No significant clusters									
Main effect sense	1	6	Supp_Motor_Area_L	−4	8	54	1389	0.000	7.71	0.000
		n/a	n/a	−10	20	22			4.44	0.146
		n/a	Cingulum_Mid_L	−10	16	32			4.12	0.387
	2	7	Occipital_Mid_L	−26	−60	42	5208	0.000	6.63	0.000
		2	Parietal_Inf_L	−46	−38	50			6.07	0.000
		37	Fusiform_L	−42	−60	−12			5.95	0.000
	3	6	Precentral_L	−46	−2	34	3488	0.000	6.53	0.000
		6	Precentral_L	−46	0	42			6.23	0.000
		44	Frontal_Inf_Oper_L	−46	8	22			6.18	0.000
	4	n/a	Thalamus_L	−10	−14	10	983	0.000	5.33	0.004
		n/a	Thalamus_L	−4	−20	12			4.99	0.018
		n/a	Thalamus_R	4	−18	6			4.61	0.079
	5	19	Occipital_Inf_R	44	−76	−2	2342	0.000	4.90	0.026
		18	Vermis_6	3	−72	−8			4.86	0.030
		7	Parietal_Sup_R	26	−58	54			4.74	0.048
	6	17	Calcarine_L	−8	−92	6	378	0.000	4.08	0.423
		19	Calcarine_L	−20	−66	6			3.97	0.556
		17	Calcarine_L	−12	−82	6			3.91	0.617
Main effect actors	1	7	Parietal_Sup_L	−28	−60	46	374	0.002	4.96	0.020
	2	44	Frontal_Inf_Oper_R	42	10	28	471	0.000	4.89	0.027
		45	Frontal_Inf_Tri_R	42	24	22			4.01	0.500
		45	Frontal_Inf_Tri_R	46	34	16			3.97	0.548
	3	37	Fusiform_R	40	−44	−18	396	0.001	4.88	0.028
		37	Temporal_Inf_R	42	−58	−10			4.76	0.045
	4	44	Frontal_Inf_Oper_L	−36	8	28	252	0.012	4.33	0.209
		n/a	Frontal_Inf_Tri_L	−38	20	24			3.66	0.875
		45	Frontal_Inf_Tri_L	−42	28	18			3.44	0.979
	5	37	Ternporal_Mid_R	36	−58	14	332	0.003	4.11	0.395
		40	n/a	30	−56	34			4.04	0.469
		19	n/a	32	−64	28			3.94	0.587
Main effect reference	1	37	Ternporal_Mid_R	54	−58	6	21483	0.000	12.62	0.000
		37	Ternporal_Mid_R	48	−64	12			11.49	0.000
		19	Occipital_Mid_L	−48	−74	4			11.27	0.000
	2	6	Frontal_Sup_L	−26	−6	55	851	0.000	5.97	0.000
	3	20	Ternporal_Mid_R	50	−10	−18	307	0.005	5.08	0.012
		20	Fusiform_R	42	−16	−20			3.62	0.903
	4	6	Frontal_Sup_R	30	−6	60	598	0.000	5.00	0.017
		n/a	Frontal_Mid_R	24	12	42			4.56	0.093
		6	Precentral_R	40	0	46			4.01	0.504
	5	6	Supp_Motor_Area_L	−4	10	56	236	0.016	4.76	0.045
		6	Frontal_Sup_L	−14	10	50			3.83	0.716
	6	27	Lingual_R	6	−34	−4	479	0.000	4.64	0.072
		27	n/a	−8	−28	−4			4.31	0.222
		27	Thalamus_L	−16	−30	6			3.84	0.702
	7	44	Precentral_L	−44	6	20	299	0.005	4.42	0.153
		6	Precentral_L	−42	−4	32			4.27	0.253

Listed are the MNI-coordinates for the first three local maxima for each significant cluster in the relevant comparisons ($p < 0.05$ corrected cluster-level, threshold $p < 0.001$ uncorrected voxelwise). Anatomical labels are derived from the Automatic Anatomical Labeling atlas (Tzourio-Mazoyer et al., 2002) and from Brodmann's atlas. Cluster-level statistics are listed for each cluster, voxel-level statistics also for local maxima.

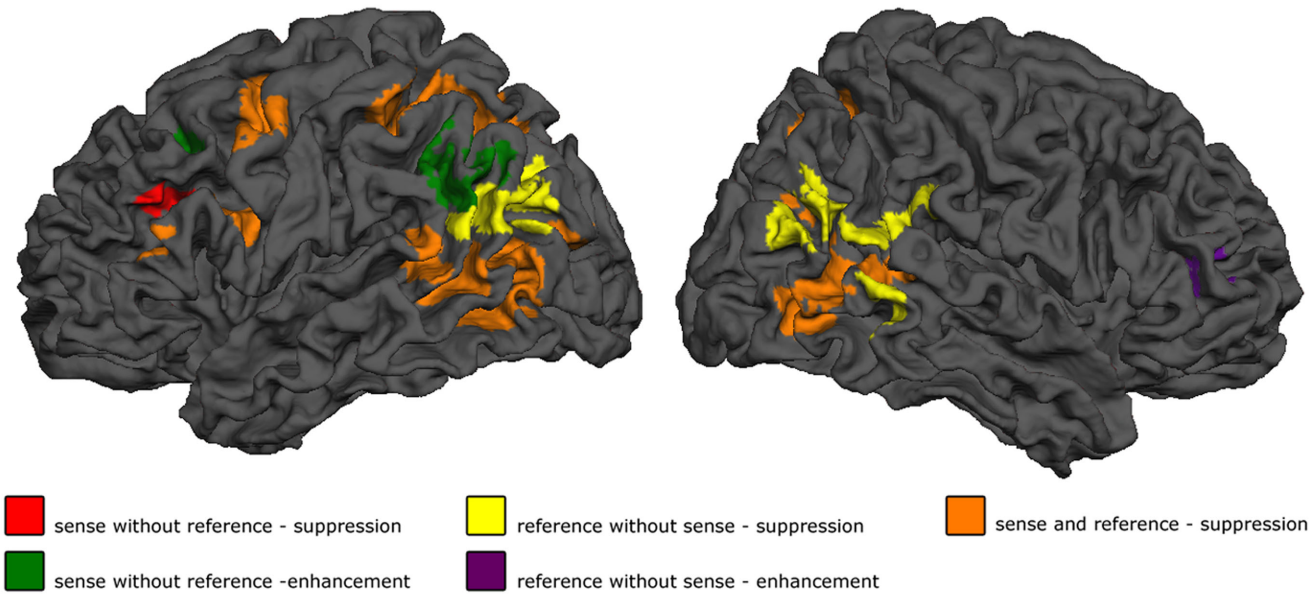


FIGURE 3 | Results. Shown are the effect of sense masked exclusively for reference, for reference masked exclusively for sense, and the conjunction of the two. $p < 0.05$ cluster-level FWE corrected for simple effects, maps thresholded at $p < 0.001$ voxelwise uncorrected, exclusive masks thresholded at $p < 0.20$ voxelwise uncorrected.

information on, for instance, space, time, intentionality, causation and protagonists (Zwaan and Radvansky, 1998). These are integrated and updated over several sentences and also contain all inferences that were not explicitly stated but are necessary for comprehension (Zwaan and Radvansky, 1998). The difference between reference as discussed above and situation models is that the latter pertain to the integration of referents of several utterances into one mental model and also contain unstated information, arrived at through inferences. A similar distinction is likely in production: the situation model may contain any information that the speaker knows pertains to the situation, but that he does not mention. Right inferior frontal gyrus has repeatedly been found to be involved in generating inferences (Mason and Just, 2004; Kuperberg et al., 2006). The first presentation of a referent may therefore induce the start of situation model construction. This same area did not show enhancement in the control experiment, supporting the idea that the process in which this region is involved is language-related. We do not currently have an explanation for the enhancement effect found in right middle temporal gyrus in the control experiment.

The second main step in semantic encoding is to map the representation that we want to talk about onto a linguistic structure that can be syntactically encoded – the actual process of encoding. This would presumably involve areas sensitive to both reference and sense, interfacing between the mental representation of the situation that will be described and the linguistic representation describing it. What is perhaps most striking about our data, is the great extent to which these two processes are neurally intertwined: bilateral posterior middle temporal gyri (BA 37), superior parietal areas (BA 7), precentral gyrus (BA 6) and LIFG (BA 44/6) all show largely overlapping suppression effects for reference and sense. Our data show that large parts of the language network are involved in processing reference, and that reference therefore presumably is important throughout much of the task of building an utterance. But what is the contribution of all these areas to semantic encoding? Due to the proximity of areas coding the perceptual representation of the referent and some of the areas involved in processing both reference and sense, we hypothesize that the bilateral temporal areas sensitive to reference and sense are primarily involved in mapping one onto the other. Such mapping requires the retrieval of the relevant lexical items from the mental lexicon, which indeed has often been postulated to involve the posterior middle temporal gyrus (Hagoort, 2005; Jung-Beeman, 2005). The bilateral superior parietal lobes also showed suppression to the repetition of both reference and sense. These parietal areas have previously been found involved in studies investigating linguistic inference (Nieuwland et al., 2007; Monti et al., 2009). In the sense/reference fMRI study discussed in the introduction, the parietal areas were more strongly activated for both referentially ambiguous and anomalous conditions compared to coherent conditions, but this effect was more pronounced for the ambiguous condition (Nieuwland et al., 2007). In a study on linguistic and logical inference, this area was found to be common to both types of inference compared to detection of grammatical violations (Monti et al., 2009). Our suppression effect in this area may reflect that in a situation where sense, reference, or both are repeated, less inferences are required than in a situation where that is not the case. The superior LIFG (BA6) also showed suppression both to repetition of sense and of reference. On the hypothesis that IFG is involved in unifying different elements into a coherent representation (Hagoort, 2005), this means that the reference of an utterance is also kept active in the working space of language. The fact that

Table 3 | Results from control experiment.

Effect	Cluster	BA	Anatomical label	Global and local maxima			Cluster-level		Voxel-level	
				x	*y*	*z*	*K*	*p*	T	*p* (FWE)
Main effect reference suppression	1	39	Temporal_Mid_L	−44	−64	16	885	0.000	6.42	0.000
		37	Temporal_Mid_L	−54	−54	4			4.97	0.015
		19	Occipital_Mid_L	−34	−76	26			3.26	0.996
	2	39	Temporal_Mid_R	40	−62	18	1379	0.000	5.66	0.001
		21	Temporal_Mid_R	46	−56	12			5.34	0.009
		37	Temporal_Mid_R	50	−66	6			5.31	0.009
Main effect reference enhancement	1	21	Termporal_Mid_R	56	−24	−8	275	0.010	5.22	0.005
Reference masked for sense	1	39	Temporal_Mid_L	−44	−64	16	695	0.000	6.42	0.000
		37	Temporal_Mid_L	−54	−54	4			4.97	0.015
		39	Occipital_Mid_L	−34	−76	26			3.26	0.996
	2	39	Temporal_Mid_R	40	−62	18	1313	0.000	5.66	0.001
		39	Temporal_Mid_R	46	−56	12			5.34	0.009
		37	Temporal_Mid_R	50	−66	6			5.31	0.009
Conjunction reference speaking and reference control	1	39	Temporal_Mid_L	−44	−64	16	885	0.000	6.42	0.000
		37	Temporal_Mid_L	−54	−54	4			4.97	0.015
		19	Occipital_Mid_L	−34	−76	26			3.26	0.996
	2	39	Temporal_Mid_R	40	−62	18	1346	0.000	5.66	0.001
		21	Temporal_Mid_R	46	−56	12			5.34	0.003
		37	Temporal_Mid_R	50	−66	6			5.31	0.003
Sense	No significant clusters									
Actors	No significant clusters									
Words	No significant clusters									

Listed are the MNI-coordinates for the first three local maxima for each significant cluster in the relevant comparisons ($p < 0.05$ corrected cluster-level, threshold $p < 0.001$ uncorrected voxelwise). Anatomical labels are derived from the Automatic Anatomical Labeling atlas (Tzourio-Mazoyer et al., 2002) and from Brodmann's atlas. Cluster-level statistics are listed for each cluster, voxel-level statistics also for local maxima.

none of the regions outlined above are sensitive to any of our factors in the control experiment further indicates that the process they are involved in is linguistic in nature.

The output of semantic encoding is the sense. One area showed a repetition suppression effect for sense but not reference: the left inferior IFG (BA 45). The final, linguistic, sense is apparently assembled in LIFG. This effect may, however, also be partly due to the repetition of the exact sentence, therefore by repetition of not just semantic but also both syntactic and phonological sequencing processes, which are related to actual speech output and are not part of the sense. In fact, the focus of the effect, lying at the heart of the part of LIFG most often found involved in syntactic processing (Bookheimer, 2002), suggests just that. Ventral LIFG, most commonly known to be involved in meaning processing (Bookheimer, 2002), remains sensitive to reference throughout.

Repetition of sense also elicits enhancement in two areas. The exact same left hemispheric frontal and parietal areas here showing repetition enhancement for sense have previously been found to be involved in semantic inhibition (Hoenig and Scheef, 2009), that is, inhibition of contextually inappropriate meanings. In the present paradigm, each word (MAN, BOY, WOMAN, GIRL) has two prominent possible referents. One of them has to be suppressed in mapping the intended referent onto the sense. While this would seem harder in the case where sense is not repeated (and therefore elicit suppression instead of enhancement upon repetition), this seeming incongruity can be readily explained: the BOLD-response in both areas shows consistent deactivation in any of the conditions compared to an implicit baseline. The deactivations are less strong in the conditions with repeated sense, than those where sense is novel. This mirrors activation patterns in the so-called default mode network, which shows increasing *de*activations depending on task difficulty (Greicius et al., 2003). Both areas have been shown to be part of the default mode network.

In sum, our data suggest that the bilateral temporo-parietal-occipital junctions are involved in constructing a mental representation of a percept (the reference), that the bilateral posterior middle temporal gyri map this representation onto lexical items that can be expressed, and that the final sense is unified in left inferior frontal gyrus – this can then serve as input to both syntactic and phonological encoding which also involve left inferior frontal gyrus.

Some caveats are in order: in operationalizing reference and sense for the purpose of this study, we have made some decisions that limit the generalizability of our findings. Most notably, our experiments concern visual representations of concrete events.

As we have stressed above, we consider referents to be mental representations in our mind. These mental representations are likely to differ depending on the material underlying them. They will likely be different for auditory and visual objects, for events involving people and for non-human objects, for concrete objects and for abstract concepts. But that is precisely the point: our brains need to convert non-linguistic mental (i.e., conceptual) representations, whatever they are "made of" into language. Therefore, while we believe our findings concerning sense, and the mapping of reference onto sense will at least to a large extent hold irrespective of the underlying reference, what brain areas are involved in processing reference alone will depend on the specifics of the mental representation involved.

Another constraint concerns our task. We had participants describe a long list of pictures. If these subsequent sentences were to be perceived as part of an ongoing discourse, then some unnatural situations would arise: normally, we would avoid repeating the same sentence twice in a row, let alone while using it to refer to different things. Our behavioral data, however, provide an indication that participants were not too affected by such concerns. First, the instructions specified that they had to name the people, the colors, and the action (which was given by the verb presented prior to the picture). Though this precluded using pronouns, this did not prevent participants from adding specifications such as "the other," "again," "now," etc., to specify the relation between pictures. No participants chose to do so. Second, if repeating the sentence were more difficult than not repeating it, we should have seen an inhibitory effect of priming. While we did see this in the speaking times, we did not in the planning times, and the total time taken to compete an utterance was shorter for the primed than the unprimed conditions. These are indications that our participants were happy to consider every trial an independent unit. We believe that single sentence processing is conceptually the same as discourse processing, but on a smaller scale. Therefore, we predict our general findings would hold for more natural processing of language in context.

To conclude, our data confirm that the theoretical distinction between reference and sense is psychologically real, both in terms of behavior and of neuroanatomy. The behavioral data shows that priming of both processes can affect the ease of production. The fMRI data shows that indeed some brain regions are selectively affected by one of these computations. However, the neuronal infrastructure underlying the computation of reference and sense is largely shared in the brain. This indicates that processing reference and sense is highly interactive throughout the language system.

ACKNOWLEDGMENTS

This research was funded by the NWO Spinoza prize awarded to Peter Hagoort. We wish to thank Geoffrey Brookshire and Josje Verhagen for their assistance in running the control experiment.

REFERENCES

Bookheimer, S. (2002). Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annu. Rev. Neurosci.* 25, 151–188.

Buur, P. F., Poser, B. A., and Norris, D. G. (2009). A dual echo approach to removing motion artefacts in fMRI time series. *NMR Biomed.* 22, 551–560.

Conrad, N., Giabbiconi, C.-M., Müller, M. M., and Gruber, T. (2007). Neuronal correlates of repetition priming of frequently presented objects: insights from induced gamma band responses. *Neurosci. Lett.* 429, 126–130.

Frege, G. (1892). "On sense and nominatum," in *The Philosophy of Language*, 2nd Edn, ed. A. P. Martinich (New York: Oxford University Press), 190–192.

Friston, K. J., Holmes, A., Worsley, K., Poline, J.-B., Frith, C., and Frackowiak, R. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210.

Friston, K. J., Penny, W. D., and Glaser, D. E. (2005). Conjunction revisited. *Neuroimage* 25, 661–667.

Gagnepain, P., Chetelat, G., Landeau, B., Dayan, J., Eustache, F., and Lebreton, K. (2008). Spoken word memory traces within the human auditory cortex revealed by repetition priming and functional magnetic resonance imaging. *J. Neurosci.* 28, 5281–5289.

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 100, 253–258.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. (2002). Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magn. Reson. Med.* 47, 1202–1210.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci. (Regul. Ed.)* 9, 416–423.

Hasson, U., Nusbaum, H. C., and Small, S. L. (2007). Brain networks subserving the extraction of sentence information and its encoding to memory. *Cereb. Cortex* 17, 2899–2913.

Henson, R., Shallice, T., and Dolan, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science* 287, 1269–1272.

Hoenig, K., and Scheef, L. (2009). Neural correlates of semantic ambiguity processing during context verification. *Neuroimage* 45, 1009–1019.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford: Oxford University Press.

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends Cogn. Sci. (Regul. Ed.)* 9, 512–518.

Kan, I., and Thompson-Schill, S. (2004). Effect of name agreement on prefrontal activity during overt and covert picture naming. *Cogn. Affect. Behav. Neurosci.* 4, 43–57.

Kuperberg, G. R., Lakshmanan, B. M., Caplan, D. N., and Holcomb, P. J. (2006). Making sense of discourse: an fMRI study of causal inferencing across sentences. *Neuroimage* 33, 343–361.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation.* Cambridge, MA: The MIT Press.

Mason, R. A., and Just, M. A. (2004). How the brain processes causal inferences in text. *Psychol. Sci.* 15, 1–7.

Menenti, L., Gierhan, S. M. E., Segaert, K., and Hagoort, P. (2011). Shared Language. *Psychol. Sci.* 22, 1174–1182.

Menenti, L., Petersson, K. M., Scheeringa, R., and Hagoort, P. (2009). When elephants fly: differential sensitivity of right and left inferior frontal gyri to discourse and world knowledge. *J. Cogn. Neurosci.* 21, 2358–2368.

Monti, M. M., Parsons, L. M., and Osherson, D. N. (2009). The boundaries of language and thought in deductive inference. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12554–12559.

Nieuwland, M. S., Petersson, K. M., and Van Berkum, J. J. A. (2007). On sense and reference: examining the functional neuroanatomy of referential processing. *Neuroimage* 37, 993–1004.

Parker Jones, O. I., Green, D. W., Grogan, A., Pliatsikas, C., Filippopolitis, K., Ali, N., Lee, H. L., Ramsden, S., Gazarian, K., Prejawa, S., Seghier, M. L., and Price, C. J. (2011). Where, when and why brain activation differs for bilinguals and monolinguals during picture naming and reading aloud. *Cereb. Cortex.* doi: 10.1093/cercor/bhr16/. [Epub ahead of print].

Poser, B. A., Versluis, M. J., Hoogduin, J. M., and Norris, D. G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: parallel-acquired inhomogeneity desensitized fMRI. *Magn. Reson. Med.* 55, 1227–1235.

Tesink, C. M. J. Y., Petersson, K. M., Van Berkum, J. J. A., Van Den Brink, D. L., Buitelaar, J. K., and Hagoort, P. (2009). Unification of speaker and meaning in language comprehension: an fMRI study. *J. Cogn. Neurosci.* 21, 2085–2099.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289.

Zwaan, R. A., and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychol. Bull.* 123, 162–185.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 July 2011; accepted: 06 December 2011; published online: 18 January 2012.

Citation: Menenti L, Petersson KM and Hagoort P (2012) From reference to sense: how the brain encodes meaning for speaking. Front. Psychology ***2****:384. doi: 10.3389/fpsyg.2011.00384*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 March 2012
doi: 10.3389/fpsyg.2012.00084

A network model of observation and imitation of speech

Nira Mashal*[1,2], *Ana Solodkin*[1,3,4], *Anthony Steven Dick*[1,5], *E. Elinor Chen*[1,3] and *Steven L. Small*[1,4]

[1] Department of Neurology, The University of Chicago, Chicago, IL, USA
[2] School of Education, Bar-Ilan University, Ramat Gan, Israel
[3] Department of Anatomy and Neurobiology, University of California, Irvine, Irvine, CA, USA
[4] Department of Neurology, University of California Irvine, Irvine, CA, USA
[5] Department of Psychology, Florida International University, Miami, FL, USA

Edited by:
Yury Y. Shtyrov, Medical Research Council, UK

Reviewed by:
Max Garagnani, MRC Cognition and Brain Sciences Unit, UK
Alessandro D'Ausilio, Istituto Italiano di Tecnologia, Italy

***Correspondence:**
Steven L. Small, Department of Neurology, University of California Irvine, Irvine, CA 92697, USA.
e-mail: small@uci.edu

Much evidence has now accumulated demonstrating and quantifying the extent of shared regional brain activation for observation and execution of speech. However, the nature of the actual networks that implement these functions, i.e., both the brain regions and the connections among them, and the similarities and differences across these networks has not been elucidated. The current study aims to characterize formally a network for observation and imitation of syllables in the healthy adult brain and to compare their structure and effective connectivity. Eleven healthy participants observed or imitated audiovisual syllables spoken by a human actor. We constructed four structural equation models to characterize the networks for observation and imitation in each of the two hemispheres. Our results show that the network models for observation and imitation comprise the same essential structure but differ in important ways from each other (in both hemispheres) based on connectivity. In particular, our results show that the connections from posterior superior temporal gyrus and sulcus to ventral premotor, ventral premotor to dorsal premotor, and dorsal premotor to primary motor cortex in the left hemisphere are stronger during imitation than during observation. The first two connections are implicated in a putative dorsal stream of speech perception, thought to involve translating auditory speech signals into motor representations. Thus, the current results suggest that flow of information during imitation, starting at the posterior superior temporal cortex and ending in the motor cortex, enhances input to the motor cortex in the service of speech execution.

Keywords: speech, language, mirror neuron, structural equation modeling, effective connectivity, action observation, ventral premotor cortex, brain imaging

INTRODUCTION

In everyday communication, auditory speech is accompanied by visual information from the speaker, including movements of the lips, mouth, tongue, and hands. Observing these motor actions improves speech perception, particularly under noisy conditions (MacLeod and Summerfield, 1987) or when the auditory signal is degraded (Sumby and Pollack, 1954; Ross et al., 2007). One putative neural mechanism postulated to account for this phenomenon is *observation–execution matching*, whereby observed actions (e.g., oral motor actions) are matched by the perceiver to a repertoire of previously executed actions (i.e., previous speech). Support for this matching hypothesis comes from recent studies showing that the brain areas active during action observation and action execution contain many shared components, and that such overlap exists for movements of the finger, hand, and arm (e.g., Tanaka and Inui, 2002; Buccino et al., 2004b; Molnar-Szakacs et al., 2005), as well as those of the mouth and lips during speech (Fadiga et al., 1999; Wilson et al., 2004; Skipper et al., 2005, 2007; D'Ausilio et al., 2011). Although these previous studies demonstrate both commonalties and differences in regional brain activation for observation and execution, they do not characterize the networks that implement these functions in terms of effective connectivity, i.e., the functional influence of one region over those with which it is anatomically connected. With such network descriptions, as we elucidate here, it is possible to show the quality and degree to which functional brain circuits for observation and execution are intertwined, and thus to test the degree of functional overlap (or lack thereof) related to the interactions established by the activated brain regions.

Studies aiming to characterize the neural mechanisms for observation and imitation of speech have used advanced brain imaging techniques and have shown that some similar brain regions are activated during the two tasks, particularly in motor regions involved in speech [i.e., ventral premotor cortex (vPM) and adjacent pars opercularis of the inferior frontal gyrus]. Although the pars opercularis has been traditionally thought to be critical for speech production (Geschwind, 1970; Ojemann et al., 1989), an increasing number of studies have shown that the underlying implementation of this function may be integrated in a multimodal fashion with visual (MacSweeney et al., 2000; Hasson et al., 2007) and audiovisual speech perception (Skipper et al., 2005, 2007). For example, silent lip-reading increases brain activity bilaterally in the premotor cortex and Broca's area (particularly pars opercularis and its homolog; MacSweeney et al., 2000), and activation in left pars opercularis is associated with individual differences in the integration of visual and auditory speech information

(Hasson et al., 2007). In macaque, related areas appear to critical for integration of parietal sensory–motor signals with higher-order information originating from multiple frontal areas, with information shared across adjacent areas (Gerbella et al., 2011).

Both passive listening to monosyllables and production of the same syllables leads to overlapping activation in a superior portion of the vPM (Wilson et al., 2004). The time course of activation on a related task – observing and imitating lip forms – successively incorporates the occipital cortex, superior temporal region, inferior parietal lobule, inferior frontal, and ultimately the primary motor cortex, with stronger activation during imitation than observation (Nishitani and Hari, 2002). Using audiovisual stimuli, we previously showed observation/execution overlap in posterior superior temporal cortices, inferior parietal areas, pars opercularis, premotor cortices, primary motor cortex, subcentral gyrus and sulcus, insula, and cerebellum (Skipper et al., 2007). Overall, a number of studies have reported engagement of speech-motor regions in visual (MacSweeney et al., 2000; Nishitani and Hari, 2002), auditory (Fadiga et al., 2002; Wilson et al., 2004; Tettamanti et al., 2005; Mottonen and Watkins, 2009; Sato et al., 2009; D'Ausilio et al., 2011; Tremblay et al., 2011), and audiovisual speech perception (Campbell et al., 2001; Fadiga et al., 2002; Calvert and Campbell, 2003; Paulesu et al., 2003; Watkins et al., 2003; Skipper et al., 2005, 2006, 2007).

The consistent activation of the pars opercularis, inferior parietal lobule, and vPM in studies of speech perception and imitation is predicted by several related accounts of audiovisual speech perception and production, and the relation between them. One set of accounts has emphasized the contribution of motor cortex to speech perception during audiovisual language comprehension (see Schwartz et al., 2012 for review). An influential perspective from this vantage point argues that motor cortex activation in speech perception is the product of "direct matching" of a perceived action with the observer's previous motor experience with that action (Rizzolatti et al., 2001). This view further hypothesizes that such matching is accomplished, at least in part, by a special class of neurons, called "mirror neurons." Mirror neurons are sensory–motor neurons, originally characterized from recordings in area F5 of the vPM of the macaque brain, that discharge during both observation and execution of the same goal-oriented actions (Fadiga et al., 1995; Strafella and Paus, 2000; Rizzolatti et al., 2001). Mirror neurons have also been identified in the rostral part of the inferior parietal cortex (areas PF and PFG) in macaque (Fogassi et al., 2005; Fabbri-Destro and Rizzolatti, 2008; Rozzi et al., 2008; for review, see Cattaneo and Rizzolatti, 2009). Mirror neurons have been found in the macaque for both oral actions and manual actions, and human imaging studies have demonstrated task-dependent functional brain activation to observation and execution that fits this pattern and suggests that mirror neurons may also exist in the human (Buccino et al., 2001; Grezes et al., 2003; for review, see Buccino et al., 2004a). Within F5, the mirror neurons are located primarily in the caudal sector in the cortical convexity of F5 (area F5c).

Visual action information from STS appears to take two different pathways to the frontal lobe, with distinct projections first to the parietal lobe and then to areas F5c and F5ab of the inferior frontal lobe. One route begins in the upper bank of the STS, and projects to PF/PFG in the inferior parietal region (Kurata, 1991; Rizzolatti and Fadiga, 1998; Nelissen et al., 2011), which corresponds roughly to the human supramarginal gyrus, and then projects to premotor area F5c. This pathway appears to emphasize information about the agent and the intentions of the agent, and comprises the parieto-frontal mirror circuit involved in visual transformation for grasping (Jeannerod et al., 1995; Rizzolatti and Fadiga, 1998). The other pathway begins on the lower bank of STS, and connects to the frontal region F5ab via the IPS (Luppino et al., 1999; Borra et al., 2008; Nelissen et al., 2011), probably subregion AIP, whereas the second emphasizes information about the object.

We recently observed a related, but topographically different, organization in the human PMv, with a ventral PMv sector containing neurons with mirror properties, and a dorsal PMv sector containing neurons with canonical properties (Tremblay and Small, 2011).

It is not known if motor cortical regions are necessary for speech perception (Sato et al., 2008; D'Ausilio et al., 2009; Tremblay et al., 2011) or are facilitatory, playing a particularly important role in situations of decreased auditory efficiency (e.g., hearing loss, noisy environment; Hickok, 2009; Lotto et al., 2009). In either case, brain networks that include frontal and parietal motor cortical regions are activated during speech perception, and may represent a physiological mechanism by which brain circuits for motor execution aid in the understanding of speech. One way this could occur is by "direct matching" (Rizzolatti et al., 2001; Gallese, 2003), in which an individual recognizes speech by mapping perceptions onto motor representations using a sensory–motor circuit including posterior inferior frontal/ventral premotor, inferior parietal, and posterior superior temporal brain regions (Callan et al., 2004; Guenther, 2006; Guenther et al., 2006; Skipper et al., 2007; Dick et al., 2010).

Although these prior studies demonstrate participation of these visuo-motor regions in speech perception, there does not yet exist a characterization of the organization of these regions into an effectively connected network relating speech production with speech perception. In this paper, we describe such a network organization, and show the relation between the human effective network for observing speech (without a goal of execution) and imitating speech (observing with a goal of execution and then executing). Specifically, we present a formal structural equation (effective connectivity) model of the neural networks used for observation and imitation of audiovisual syllables in the normal state, and compare the structure and effective connectivity of observation and imitation networks in both left and right hemispheres.

In investigating these questions, we have three hypotheses. (i) First, we postulate a gross anatomical similarity between the networks for observation and imitation, i.e., optimal models of the raw imaging data can be described with a core of similar nodes (regions), since there will be overlapping regional activation during both observation and imitation. (ii) Second, we suggest that the effective connections within the network will be of approximately equal strength, particularly those with larger motor biases, such as the connection between the inferior frontal/ventral premotor regions and the inferior parietal regions. (iii) Third, we expect that the networks with the best fit to the data will differ between the left and right hemispheres for both observation and imitation, based

on the postulated left-hemispheric bias for auditory language processing (Hickok and Poeppel, 2007). Further, based on previous findings in speech perception and auditory language understanding (e.g., Mazoyer et al., 1993; Binder et al., 1997) and imitation (e.g., Saur et al., 2008), we expect stronger effective connectivity among relevant regions in the left hemisphere (LH) during imitation compared to observation since the former requires speech output (e.g., see Nishitani and Hari, 2002 for a discussion).

To test these three hypotheses, we focused on six regions that have been shown in previous studies to be involved in speech perception. These regions include (i) vPM and inferior frontal gyrus (combined region); (ii) inferior parietal lobule (including intraparietal sulcus); (iii) primary motor and sensory cortices (M1S1); (iv) dorsal premotor cortex (dPM); (v) posterior superior temporal gyrus and sulcus (combined region); and (vi) anterior superior temporal gyrus and sulcus (combined region).

MATERIALS AND METHODS

PARTICIPANTS

Eleven adults (seven females, mean age = 24 ± 5) participated. All were right handed as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971), with no history of neurological or psychiatric illness. Participants gave written informed consent and the Institutional Review Board of the Biological Sciences Division of The University of Chicago approved the study.

STIMULI AND TASK

Participants performed two tasks. In the Observation task, participants passively watched and listened to a female actress (filmed from neck up) articulating four syllables with different articulatory profiles in terms of lip and tongue movements: /pa/, /fa/, /ta/, and /tha/. In the Imitation task, participants were asked to say the syllable out loud immediately after observing the same actress. Each syllable was presented for 1.5 s. The Observation run was 6′30″ (6 minutes and 30 seconds) in duration (260 whole-brain images) and the Imitation run lasted 12′30″ (12 minutes and 30 seconds) (500 whole-brain images). Each run contained a total of 120 stimuli (30 stimuli for each syllable). In each of these runs, stimuli were presented in a randomized event-related manner with a variable interstimulus interval (ISI; minimum ISI for Observation = 0 s; minimum ISI for Imitation = 1.5 s, maximum ISI = 12 s for both runs). The ISI formed the baseline for computation of the hemodynamic response. Participants viewed the video stimuli through a mirror attached to the head coil that allowed them to see a screen at the end of the scanning bed. The audio track was simultaneously delivered to participants at 85 dB SPL via headphones containing MRI-compatible electromechanical transducers (Resonance Technologies, Inc., Northridge, CA, USA). Before the beginning of the experiment, participants were trained inside the scanner with a set of four stimuli to ensure they understood the tasks and could hear properly the voice of the actress.

IMAGING AND DATA ANALYSIS

Functional imaging was performed at 3 T (TR = 1.5 s; TE = 25 ms; FA = 77°; 29 axial slices; 5 mm × 3.75 mm × 3.75 mm voxels) on a GE Signa scanner (GE Medical Systems, Milwaukee, WI, USA) using spiral BOLD acquisition (Noll et al., 1995). A volumetric T1-weighted inversion recovery spoiled grass sequence (120 axial slices, 1.5 mm × 0.938 mm × 0.938 mm resolution) was used to acquire structural images on which anatomical landmarks could be found and functional activation maps could be superimposed.

DATA ANALYSIS

Preprocessing and identification of task-related activity

Functional image preprocessing for each participant consisted of three-dimensional motion correction using weighted least-squares alignment of three translational and three rotational parameters, as well as registration to the first non-discarded image of the first functional run, and to the anatomical volumes (Cox and Jesmanowicz, 1999)[1]. The time series were linearly detrended and despiked, the impulse response function was estimated using deconvolution, and analyzed statistically using multiple linear regression. The two principal regressors were for the Observation task and the Imitation task. Nine sources of non-specific variance were removed by regression, including six motion parameters, the signal averaged over the whole-brain, the signal averaged over the lateral ventricles, and the signal averaged over a region centered in the deep cerebral white matter. The regressors were converted to percent signal change values relative to the baseline, and significantly activated voxels were selected after correction for multiple comparisons using false discovery rate (FDR; Benjamini and Hochberg, 1995; Genovese et al., 2002) with a whole-brain alpha of $p < 0.05$.

Whole-brain group analysis of condition differences

A group analysis was conducted on the whole-brain to determine whether there was a significant group level activation relative to a resting baseline, and to compare condition differences at the group level. We conducted one-sample t tests to assess activation relative to zero, and dependent paired-sample t tests to assess condition differences. These were computed on a voxel-wise basis using the normalized regression coefficients as the dependent variable. To control for multiple comparisons, we used the FDR procedure ($p < 0.05$).

Network analysis using structural equation modeling

The primary analysis was a network analysis using SEM (McIntosh, 2004), which was performed using AMOS software (Arbuckle, 1989), which can be used to model fMRI data from both block and event-related designs (Gates et al., 2011). We first specified a theoretical anatomical model, which consisted of the regions comprising the nodes of the network, and the directional connections (i.e., paths) among them. Our hypotheses focused on six anatomical regions, identified on each individual participant. The regions of the model, which are specified further in **Table 1**, included M1S1, dPM, vPM including *pars opercularis* of the inferior frontal gyrus and the inferior portion of the precentral sulcus and gyrus, inferior parietal lobule (IP) including the intraparietal sulcus, posterior superior temporal gyrus and sulcus (pST), and anterior superior temporal gyrus and sulcus (aST). Connections were specified with reference to known macaque anatomical connectivity (e.g.,

[1] http://afni.nimh.nih.gov/afni/

Table 1 | Anatomical description of the cortical regions of interest.

ROI	Anatomical structure	Brodmann's area	Delimiting landmarks
IFGOp/PMv	*Pars opercularis* of the inferior frontal gyrus, inferior precentral sulcus, and inferior precentral gyrus	6, 44	A = anterior vertical ramus of the sylvian fissure P = central sulcus S = inferior frontal sulcus, extending a horizontal plane posteriorly across the precentral gyrus I = anterior horizontal ramus of the sylvian fissure to the border with insular cortex
PMd	*Pars opercularis* of the inferior frontal gyrus, inferior precentral sulcus, and inferior precentral gyrus	6	A = vertical plane through the anterior commissure P = central sulcus S = medial surface of the hemisphere I = inferior frontal sulcus, extending a horizontal plane posteriorly across the precentral gyrus
IP	Supramarginal gyrus; angular gyrus; intraparietal sulcus	39, 40	A = postcentral sulcus P = sulcus intermedius secundus S = superior parietal gyrus I = horizontal posterior segment of the superior temporal sulcus
STa	Anterior portion of the superior temporal gyrus, superior temporal sulcus, and planum polare	22	A = inferior circular sulcus of insula P = a vertical plane drawn from the anterior extent of the transverse temporal gyrus S = anterior horizontal ramus of the sylvian fissure I = middle temporal gyrus
STp	Posterior portion of the superior temporal gyrus, superior temporal sulcus, and planum temporale	22, 42	A = a vertical plane drawn from the anterior extent of the transverse temporal gyrus P = angular gyrus S = supramarginal gyrus I = middle temporal gyrus
	Central sulcus; postcentral gyrus	1, 2, 3, 4	A = precentral gyrus P = postcentral sulcus S = medial surface of the hemisphere I = parietal operculum

A, anterior; P, posterior; S, superior; I, inferior.

Petrides and Pandya, 1984; Matelli et al., 1986; Seltzer and Pandya, 1994; Rizzolatti and Matelli, 2003; Schmahmann et al., 2007)

Definition of these regions on each individual participant was obtained using the automated parcelation procedure in Freesurfer[2]. Cortical surfaces were inflated (Fischl et al., 1999a) and registered to a template of average curvature (Fischl et al., 1999b). The surface representations of each hemisphere of each participant were then automatically parcelated into regions (Fischl et al., 2004). Small modifications to this parcelation were made manually (see **Table 1** for anatomical definition).

For SEM, we first re-sampled the (rapid event-related) time series to enable assessment of variability and thus quantification of goodness of fit. We first obtained time series from the peak voxel in each ROI (voxel associated with the highest t value from all active voxels; corrected FDR $p < 0.05$). The peak voxel approach was chosen because it has been shown empirically in comparison with other approaches to result in robust models across individual participants contributing to a group model (Walsh et al., 2008). We then re-sampled these time series (260 and 500 time points for the Observation and Imitation conditions, respectively) down to 78 in the LH and 77 points in the RH using a locally weighted scatterplot smoothing (LOESS) method. In this method, each re-sampled data point is estimated with a weighted least-squares function, giving greater weight to actual time points near the point being estimated, and less weight to points farther away (Cleveland and Devlin, 1988). Non-significant Box's M tests indicated no differences in the variance–covariance structure of the re-sampled and original data. The SEM analysis was conducted on these re-sampled time series.

To specify a theoretical model constrained by known anatomy, and to determine whether it was able to reproduce the observed data, we used maximum likelihood estimation. We first estimated the path coefficients based on examination of the interregional correlations, which were used as starting values to facilitate maximum likelihood estimation (McIntosh and Gonzalez-Lima, 1994). We assessed the difference between the predicted and the observed solution using the stacked model (multiple group) approach (Gonzalez-Lima and McIntosh, 1994; McIntosh and Gonzalez-Lima, 1994; McIntosh et al., 1994). If the χ^2 statistic characterizing

[2]http://surfer.nmr.mgh.harvard.edu

the difference between the models in not significant, then the null hypotheses (i.e., that there is no difference between the predicted and the observed data) should be retained, and the model represents a good fit. Note that in cases where two models have different degrees of freedom, missing nodes are included with random–constant time series and its connections are added to the less specified model with connection strength of zero to permit comparison (Solodkin et al., 2004).

RESULTS

WHOLE-BRAIN ANALYSIS: ACTIVATION COMPARED TO RESTING BASELINE AND ACROSS CONDITIONS

Patterns of activation in Imitation and Observation conditions were quite similar, with activation in the occipital cortex, anterior and posterior superior temporal regions, inferior frontal gyrus, and primary sensory–motor cortex bilaterally. All activations were of higher volume and intensity during Imitation compared to Observation (see **Table 2** for the quantitative data). Activation during Imitation but not Observation extended to anterior parts of the IFG (i.e., pars orbitalis) bilaterally. The activation profile from a representative participant is shown in **Figure 1**.

STRUCTURAL EQUATION MODELS: MODELS OF OBSERVATION AND IMITATION IN THE LEFT HEMISPHERE

The predicted model for the LH fit the data for both Observation and Imitation (for Observation: $\chi^2 = 1.8$, df = 1, $p = 0.18$; for Imitation $\chi^2 = 0.01$, df = 1, $p = 0.92$; see **Figure 2**). The strongest effective connections (EF > 0.4) for both Observation and Imitation models included those from pST to IP (0.60, 0.72, in Observation and Imitation, respectively), from IP to vPM (0.63, 0.47, respectively), from vPM to M1S1 (0.81, 0.73), and from pST to aST (0.54, 0.50, respectively).

There were also important differences based on comparison with the stacked model approach (Gonzalez-Lima and McIntosh, 1994; McIntosh and Gonzalez-Lima, 1994; McIntosh et al., 1994). Overall, the models for Observation and Imitation differed ($\chi^2 = 55.2$, df = 18, $p < 0.0001$), suggesting differences in the magnitude of some of the path coefficients of the models. Although many of the coefficients did not differ (**Figure 3**; including IP → vPM, pST → aST, pST → IP, aST → IP, IP → dPM, IP → M1S1, vPM → M1S1), connections from pST to vPM ($\chi^2 = 11.7$, df = 1, $p < 0.001$), vPM to dPM ($\chi^2 = 5.2$, df = 1, $p < 0.05$), and dPM to M1S1 ($\chi^2 = 14.0$, df = 1, $p < 0.001$) were stronger during Imitation than during Observation (**Figure 4**).

Table 2 | Average number of active voxels in each of the regions of interest (FDR corrected $p < 0.05$).

	Observation		Imitation	
Region	LH	RH	LH	RH
IP	32	36	79	79
M1S1	5	2	52	47
pST	22	53	38	67
aST	16	20	36	37
vPM	25	25	95	67
dPM	17	15	78	60

*LH, left hemisphere; RH, right hemisphere; IP, inferior parietal lobule; M1S1, primary motor/somatosensory cortex; pST, posterior superior temporal gyrus and sulcus; aST, anterior superior temporal gyrus and sulcus; vPM, ventral premotor cortex; dPM, dorsal premotor cortex. Anatomical definition of the regions is provided in **Table 1**.*

STRUCTURAL EQUATION MODELS: MODELS OF OBSERVATION AND IMITATION IN THE RIGHT HEMISPHERE

Connectivity models with similar nodes characterized both the Observation and Imitation conditions (for Observation: $\chi^2 = 2.5$, df = 1, $p = 0.11$; for Imitation: $\chi^2 = 3.1$, df = 2, $p = 0.21$; **Figure 2**). As in the case of the LH, there were differences in the magnitude of the path coefficients across conditions ($\chi^2 = 174.3$, df = 17, $p < 0.001$). As can be seen from **Figure 3**, some of the connections are different during Imitation from Observation: IP → vPM ($\chi^2 = 15.4$, df = 1, $p < 0.01$), pST → vPM ($\chi^2 = 7.4$, df = 1, $p < 0.01$), pST → M1S1 ($\chi^2 = 7.2$, df = 1, $p < 0.01$), vPM → dPM ($\chi^2 = 11.5$, df = 1, $p < 0.001$), and vPM → M1S1 ($\chi^2 = 28.5$, df = 1, $p < 0.001$). In contrast with the LH, some of the connections were different during Observation from Imitation: pST → IP ($\chi^2 = 14.4$, df = 1, $p < 0.001$), pST → dPM ($\chi^2 = 4.5$, df = 1, $p < 0.05$), and IP → M1S1 ($\chi^2 = 12.2$, df = 1, $p < 0.001$).

STRUCTURAL EQUATION MODELS: MODELS OF IMITATION IN LH VS RH

The models for Imitation differed across hemispheres ($\chi^2 = 84.2$, df = 18, $p < 0.0001$). Specifically, three connections were different in the LH compared to the RH: pST → IP ($\chi^2 = 16.8$, df = 1, $p < 0.001$), vPM → M1S1 ($\chi^2 = 17.7$, df = 1, $p < 0.001$), aST → dPM ($\chi^2 = 7$, df = 1, $p < 0.05$). No connections were different in the RH from the LH.

STRUCTURAL EQUATION MODELS: MODELS OF OBSERVATION IN THE LH VS RH

The models for Observation differed across hemispheres ($\chi^2 = 50.8$, df = 16, $p < 0.001$). Specifically, three connections were different in the LH than the RH: IP → vPM ($\chi^2 = 13.8$, df = 1, $p < 0.001$), IP → dPM ($\chi^2 = 4.6$, df = 1, $p < 0.05$), and pST → M1S1 ($\chi^2 = 6.7$, df = 1, $p < 0.01$).

DISCUSSION

The present study examined three hypotheses regarding effective connectivity among brain regions important for observation and imitation of audiovisual syllables in the healthy adult. In our first hypothesis, we predicted structural similarity (i.e., similar active regions) across conditions, and our findings support this: The networks for Observation and Imitation incorporate the same nodes (brain regions). In our second hypothesis, we predicted similarity in regional interconnectivity across Observation and Imitation, and found partial support for this: while we did find considerable similarity across Imitation and Observation (e.g., see **Figure 4**), we also found several differences in connectivity in both hemispheres. Interestingly, the effective connectivity differences are not restricted to connections between historically identified "motor" areas (e.g., the connection between ventral premotor to dorsal premotor), as would be expected when motor execution is necessary for Imitation but not for Observation. While we did find this,

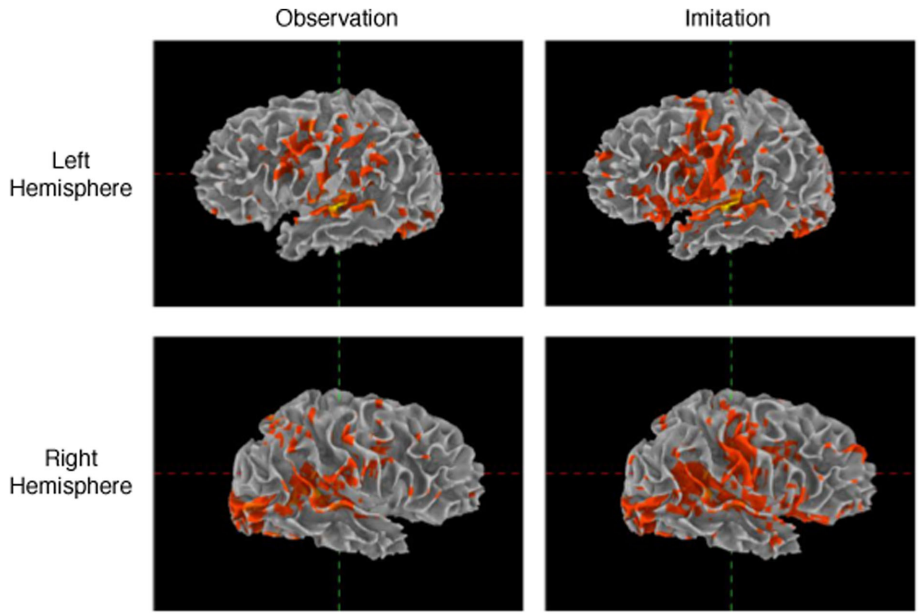


FIGURE 1 | Activation during observation and imitation. Voxels were selected using general linear model after adjusting for false positives using false discovery rate ($p < 0.05$). The figure shows the data obtained from a representative single subject.

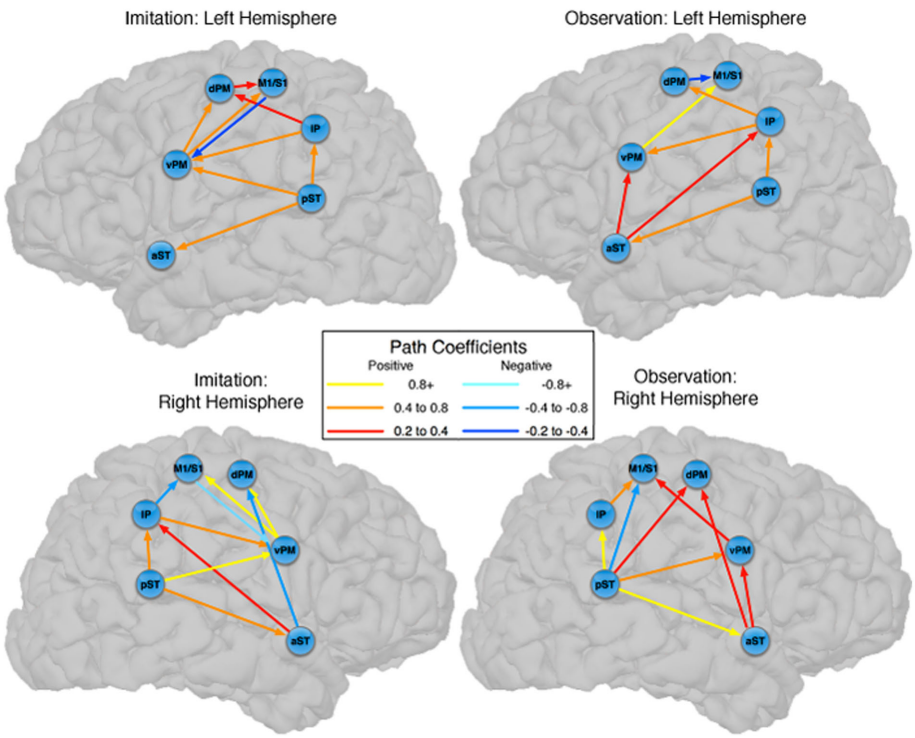


FIGURE 2 | Observation and Imitation models in both the LH and RH with connections between pST, aST, IP, vPM, dPM, and M1S1. IP, inferior parietal lobule. M1S1, primary motor/somatosensory cortex; pST, posterior superior temporal gyrus and sulcus; aST, anterior superior temporal gyrus and sulcus; vPM, ventral premotor cortex; dPM, dorsal premotor cortex; M1/S1, primary motor/somatosensory cortex. Anatomical definitions of the regions are provided in **Table 1**.

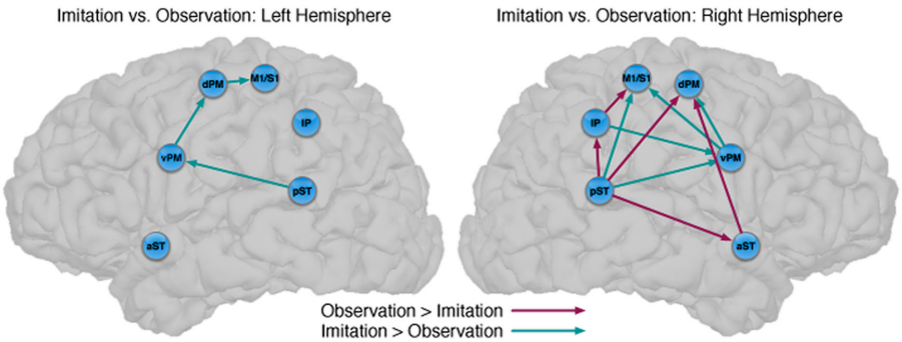


FIGURE 3 | Comparison between the models of observation and imitation in LH and in the RH. The connections between pST, aST, IP, vPM, dPM, and M1S1 showing stronger connection weights for Imitation vs Observation, and stronger connection weights for Observation vs Imitation. The flow of information in the LH might suggest a pathway to execute speech during Imitation. For a key to abbreviations, please see the legend to **Figure 2** and **Table 1**.

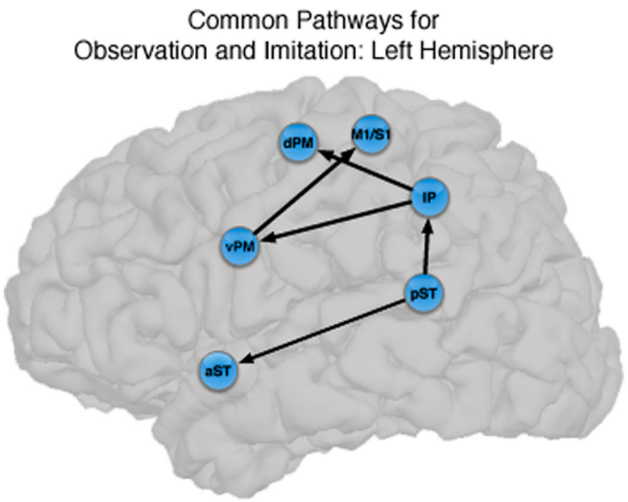


FIGURE 4 | Common pathways for observation and imitation in the LH. Observation and Imitation models in the LH with connections between pST, aST, IP, vPM, dPM, and M1S1. Black arrows show the connections that did not differ statistically across the Observation and Imitation models. For a key to abbreviations, please see the legend to **Figure 2** and **Table 1**.

we also found differences in sensory–motor interactions (e.g., the connection between posterior superior temporal region and vPM). In our third hypothesis, we predicted stronger effective connectivity during Imitation compared to Observation, particularly in the LH, since the former requires speech output and the latter does not. For the "dorsal stream" pathway connecting pST → vPM → dPM and M1/S1, we found with stronger connectivity for Imitation compared to Observation in both hemispheres. Additional differences in connectivity were found across the two conditions in the right hemisphere (RH).

It is important to note that these models reflect effective connectivity and not anatomical connectivity. Thus, whereas we show the presence of overlapping networks for Observation and Imitation, characterized by similar anatomical regions and similar statistical covariation among activity in these regions, we cannot make conclusions about brain anatomy, i.e., the white matter connections among these regions. SEM does not assess anatomical pathways directly, but rather statistical covariance in the BOLD response. Nevertheless, these networks present strong evidence on effective connectivity, which incorporates an *a priori* anatomical model (based largely on what is known about connectivity in the non-human primate), but still represents statistical covariation and not explicit anatomical evidence, supporting a human system for observation–imitation matching in speech perception.

OBSERVATION AND IMITATION IN THE LH

The results we report with respect to BOLD signal amplitude replicate previous studies in audiovisual speech perception that show brain activation in regions involved in planning and execution of speech (Calvert et al., 2000; Callan et al., 2003, 2004; Calvert and Campbell, 2003; Jones and Callan, 2003; Sekiyama et al., 2003; Wright et al., 2003; Miller and D'Esposito, 2005; Ojanen et al., 2005; Skipper et al., 2005, 2007; Pekkola et al., 2006; Pulvermuller et al., 2006; Bernstein et al., 2008). Specifically, we showed that several regions were active during both speech production and speech perception (**Table 2**; cf. (Pulvermuller et al., 2006; Skipper et al., 2007). These regions were also activated in an event-related MEG study (Nishitani and Hari, 2002), which showed temporal progression of activity for both observation and imitation (of static lip forms) from the occipital cortex to the pST, the IP, IFG, to the sensory–motor cortex (M1S1). In our current work, we elaborate on these activation studies by elucidating the functional relationships between the relevant regions, i.e., showing the basic organization of the network in terms of effective connections and relative strengths across conditions and hemispheres.

The novel contribution of the present work is a characterization of the networks for observation and imitation of dynamic speech stimuli, and we found that the functional interactions among brain regions that were active during Observation and Imitation share both similarities and differences. **Figure 4** illustrates the connections with similar strength during both conditions in the LH. The current models are consistent with the time course demonstrated by prior MEG results (Nishitani and Hari, 2002) and with previous models of effective connectivity during the perception of

intelligible speech (between superior temporal and inferior frontal regions; Leff et al., 2008) and speech production (between inferior frontal/ventral premotor regions and primary motor cortex during production; Eickhoff et al., 2009). We consider these similarities below.

The models presented here include integral connections from pST → IP, from IP → IFG/vPM, and from IFG/vPM → M1S1. Both pST and IP have been implicated in speech perception, and both are activated during acoustic and phonological analyses of speech (e.g., Binder et al., 2000; Burton et al., 2000; Wise et al., 2001). pST is activated by observation of biologically relevant movements including mouth, hands, and limb movements (Allison et al., 2000). This model represents a hierarchical network from the sensory temporal and parietal lobules, to inferior frontal and ventral premotor regions, to execution by primary motor cortex. In fact, it is quite similar, in many respects, to the results presented by Nishitani and Hari (2002) in their MEG study of observation and imitation of static speech stimuli. These authors identified a flow of information from posterior superior temporal sulcus, to inferior parietal lobule, to inferior frontal cortex, to primary motor cortex.

Our results are also consistent with those of Leff et al. (2008), who investigated word-level language comprehension, and exhaustively constructed all models of effective connectivity across the posterior superior temporal, anterior superior temporal, and inferior frontal gyrus. They found that the optimal model exhibited a "forward" architecture originating in the posterior superior temporal sulcus, with a directional projection to the anterior superior temporal sulcus, and a subsequent termination in the anterior inferior frontal gyrus. Notably, unlike the model proposed by Nishitani and Hari (2002), Leff et al.s' (2008) model of temporal–inferior frontal connectivity did not pass through the inferior parietal lobule. We found a similar pathway, in addition to the "forward" architecture from pST → IP → vPM pathway, in which there was significant directional connectivity from pST → aST. We showed that this directional pST → aST connection is present during both Observation and Imitation. This finding provides support for the notion that shared network interactions during production and perception allow for the development of and maintenance of speech representations, in particular between anterior and posterior superior temporal regions typically emphasized during speech perception and comprehension. However, the fact that we did not find strong connectivity between the aST and IFG/vPM during Imitation suggests the core interactions shared by Observation and Imitation proceed through the "dorsal" pST → IP → IFG/vPM → M1/S1 pathway identified by Nishitani and Hari (2002).

Of particular interest is that the connection strengths from IP to IFG/vPM, and from IFG/vPM to M1S1 were not significantly different across conditions. Interactions between IP and IFG/vPM have been shown to be important for speech production, as electrical stimulation of both of these structures and the fiber pathways connecting them impairs speech production (Duffau et al., 2003). Further, both of these regions are sensitive to the incongruence between visual and auditory speech information during audiovisual speech perception (Hasson et al., 2007; Bernstein et al., 2008). In conjunction with the primary sensory–motor cortex, the vPM, and posterior inferior frontal gyrus are also necessary for speech production (Ojemann et al., 1989; Duffau et al., 2003), and there is evidence that even perception of audiovisual and auditory-only speech elicits activity in both premotor and primary motor cortices (Pulvermuller et al., 2006; Skipper et al., 2007). Thus, in addition to the overlapping regional activation for observation and imitation of audiovisual speech, we show similar connectivity from IP to IFG/vPM and from IFG/vPM to M1S1 during these tasks, suggesting similar interactivity among these regions during perception and production of speech. This finding suggests that the flow of information during speech perception involves a motor execution circuit, and this motor circuit (IP – IFG/vPM – M1S1) supporting speech production relies on the relevant sensory experience.

Although there are similarities, the networks implementing perception and production of speech are dissociated by stronger effective connectivity in the LH for Imitation compared to Observation (**Figure 3**). Connections that differed included those from pST to vPM, vPM to dPM, and dPM to M1S1, all of which were stronger during Imitation than during Observation. The first two connections (pST → vPM, vPM → dPM) are implicated in Hickok and Poeppel's (2007) "dorsal stream" of speech perception. By their account, the dorsal stream helps translate auditory speech signals into motor representations in the frontal lobe, which is essential for speech development and normal speech production (Hickok and Poeppel, 2007). Our results are consistent with this view by pointing to a partially overlapping network for Observation and Imitation as part of a larger auditory–motor integration circuit. The presence of a connection from dPM to M1S1 that is stronger during Imitation than Observation represents a novel finding of potential relevance, suggesting a flow of information during Imitation from pST to vPM, vPM to dPM, and dPM to M1S1, which provides stronger input to M1S1 in triggering speech execution.

The LH models also differed across tasks by the inclusion of a negative influence from dPM to M1S1 during Observation that was positive during Imitation. Such negative influence in the motor system has been previously shown in a motor imagery task, compared to overt motor execution (Solodkin et al., 2004). In that study participants were asked to execute finger–thumb opposition movement or to imagine it kinetically (with no overt motor output). In the model that describes the execution of movement, dPM had positive influence on M1S1 whereas during kinetic imagery M1S1 received strong negative influence. This is consistent with the recent argument that action observation involves some sort of covert simulation (Lamm et al., 2007) that has similarities with kinetic motor imagery (Fadiga et al., 1999; Solodkin et al., 2004). Although the precise nature of such a mechanism remains elusive, and appears not to make use of identical circuits (Tremblay and Small, 2011), the present network models, with their shared but distinctive features, suggest a more formal notion of what such "simulation" might mean in terms of network dynamics.

Our data also support the notion that imitation of speech in the human brain involves a hierarchical flow of information from pST to IP to vPM through the "dorsal stream." As noted, Nishitani and Hari (2002) found evidence for this pathway during observation and imitation of static speech, and Iacoboni (2005) called this small circuit the "minimal neural architecture for imitation."

By this account, the STS sends a visual description of the observed action to be imitated to posterior parietal mirror neurons, then augments it with additional somatosensory input before sending to inferior frontal mirror neurons, which code for the associated goal of the action. Efferent copies of the motor commands providing the predicted sensory output are then sent to sensory cortices and comparisons are made between real and predicted sensory consequences, and corrections are made prior to execution. We have previously developed a model of speech perception based on an analogous mechanism (Skipper et al., 2006).

Thus, our data demonstrate that the core circuit underlying imitation of speech in the LH overlaps with that for observation, and that this circuit is embedded in larger networks that differ statistically. This supports the view that the core circuitry of imitation (pST, IP, vPM) in a context-dependent manner (McIntosh, 2000) depending on the nature of the actions to be imitated (Iacoboni et al., 2005).

OBSERVATION AND IMITATION IN THE RH

We have established similarities and differences for speech imitation and observation in the LH, but how is the observation of speech processed similarly or differently from the imitation of speech in the RH? We first focus on the similarities across hemispheres. For the pST → vPM → dPM component of a "dorsal" pathway, both hemispheres showed stronger connectivity during Imitation compared to Observation (turquoise in **Figure 3**). However, unlike in the LH, in the RH during Imitation the M1/S1 region was more influenced by activity in the vPM rather than the dPM. Inferior parietal → vPM connectivity was also stronger during Imitation in the RH. These results suggest strong similarities in the pathways for speech production across hemispheres, although there are differences, primarily in the interactions between vPM and inferior parietal and primary motor/somatosensory regions. Further, with the exception of some differences in premotor–motor interactions, a dorsal stream implemented through pST–vPM interactions appears to be a prominent component for both Imitation and Observation in both hemispheres.

We are not making the claim that the two hemispheres are involved in speech production in an identical manner. It is well known that LH damage leads to more severe impairments in speech production and articulation (Dronkers, 1996; Borovsky et al., 2007) and there is evidence for increasing LH involvement in speech production with development (Holland et al., 2001). However, we do emphasize that the predominant focus of the prior literature on LH involvement minimizes the involvement of the RH, it ignores the fact that different regions show different patterns of lateralization, and it does not provide a sufficient characterization of how different regions in the speech production network interact. For example, there are regional differences in the developmental trajectory of lateralization for speech production. While the left inferior frontal/vPM shows increasing lateralization with age during speech production tasks (Holland et al., 2001; Brown et al., 2005; Szaflarski et al., 2006), this pattern does not hold for posterior superior temporal and inferior parietal regions, which show a more bilateral pattern of activation (Szaflarski et al., 2006). With respect to connectivity, despite evidence for the participation of both hemispheres in speech production (Abel et al., 2011; Elmer et al., 2011; see Indefrey, 2011 for review), other effective connectivity models of speech production (e.g., Eickhoff et al., 2009), have failed to model the connectivity of RH regions. We have done so here, and have revealed interesting differences in the interactions of sensory and motor regions during speech perception and production across both hemispheres.

It is notable that the only connections in which Observation was stronger than Imitation were found in the RH, and this provides further support for the notion that speech perception also relies on the participation of the RH (McGlone, 1984; Boatman, 2004; Hickok and Poeppel, 2007). For the RH, our model exhibits a similar "forward" pST → aST architecture described by Leff et al. (2008) in their network study of speech comprehension, such that in our study for Observation compared to Imitation activation in the pST modulated activation in M1/S1 via IP, and in dPM both directly and via aST. This latter connection also mirrors the pST–aST influence from Leff and colleagues, but our results further suggest that these interactions continue to influence nodes of the network typically associated with motor output.

Our findings are also consistent with evidence from electrocortical mapping suggesting that information transfer during speech perception proceeds from the posterior superior temporal cortex in both an anterior direction (to the anterior superior temporal cortex; Leff et al., 2008) and in a posterior direction through the inferior parietal lobe (see Boatman, 2004 for review). We suggest that the modulation of premotor and motor regions via these functional paths (pST → IP → M1/S1; and pST → aST → dPM; pST → dPM) during Observation could reflect the modification of speech-motor representations through perception. That is, while to this point we have focused on action/motor influences on speech perception, there is also evidence that speech perception shapes articulatory/motor representations. This notion is more evident over the course of development, where speech perception and speech production emerge in concert over an extended period (Doupe and Kuhl, 1999; Werker and Tees, 1999), but such influences remain a part of models of adult speech perception (e.g., Guenther, 2006; Guenther et al., 2006; Schwartz et al., 2012) and have some support from functional imaging studies of the role of the pST in the acquisition and maintenance of fluent speech (Dhanjal et al., 2008).

This explanation requires further empirical investigation, and it also raises the question of why such differences during Observation and Imitation were not revealed in the LH. Functional interactions among these regions (e.g., aST → dPM) were either not significant in the LH, or did not differ significantly across conditions (see **Figure 4**). Null findings are difficult to interpret, and we cannot rule out the possibility that intermediate nodes that were not modeled have an influence on the connectivity profile of these networks. For example, although Eickhoff et al. (2009) found significant cortico-cerebellar and cortico-striatal interactions in their dynamic causal model of speech production, we did not model these interactions, and this could account for the difference in the connectivity profile across hemispheres. Alternatively, it may also reflect that the core circuit underlying imitation and observation of speech in the LH overlaps considerably not only in the structure of the functional connections, but also in the strength of the interactions.

In summary, similar, if not identical, brain networks medicate the observation and imitation of audiovisual syllables, suggesting strong overlap in the neural implementation of speech production and perception. The network for Imitation in particular appears to be mediated by the two cerebral hemispheres in similar ways. In both hemispheres during both Observation and Imitation, there is significant directional connectivity between pST → aST. However, the primary flow of audiovisual speech information involves a "dorsal" pathway proceeding from pST → IP → vPM → M1/S1, with additional modulation of M1/S1 through dPM. The regions that appear to have mirror properties in humans, IP and vPM, are functionally integrated with temporal regions involved in speech perception and motor and somatosensory regions involved in speech production, and comprise the core of this network.

ACKNOWLEDGMENTS

This study was supported by the National Institutes of Health (NIH) under grants NIH R01 DC007488 and DC03378 (to Steven L. Small), and NS-54942 (to Ana Solodkin), and by the James S. McDonnell Foundation under a grant to the Brain Network Recovery Group.

REFERENCES

Abel, S., Huber, W., Weiller, C., Amunts, K., Eickhoff, S., and Heim, S. (2011). The influence of handedness on hemispheric interaction during word production: insights from effective connectivity analysis. *Brain Connect.*

Allison, T., Puce, A., and Mccarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends Cogn. Sci. (Regul. Ed.)* 4, 267–278.

Arbuckle, J. L. (1989). AMOS: analysis of moment structures. *Am. Stat.* 43, 66–67.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Method.* 289–300.

Bernstein, L. E., Lu, Z. L., and Jiang, J. (2008). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Res.* 1242, 172–184.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., and Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cereb. Cortex* 10, 512–528.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., and Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *J. Neurosci.* 17, 353–362.

Boatman, D. (2004). Cortical bases of speech perception: evidence from functional lesion studies. *Cognition* 92, 47–65.

Borovsky, A., Saygin, A. P., Bates, E., and Dronkers, N. (2007). Lesion correlates of conversational speech production deficits. *Neuropsychologia* 45, 2525–2533.

Borra, E., Belmalih, A., Calzavara, R., Gerbella, M., Murata, A., Rozzi, S., and Luppino, G. (2008). Cortical connections of the macaque anterior intraparietal (AIP) area. *Cereb. Cortex* 18, 1094–1111.

Brown, T. T., Lugar, H. M., Coalson, R. S., Miezin, F. M., Petersen, S. E., and Schlaggar, B. L. (2005). Developmental changes in human cerebral functional organization for word generation. *Cereb. Cortex* 15, 275–290.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., and Freund, H. J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *Eur. J. Neurosci.* 13, 400–404.

Buccino, G., Binkofski, F., and Riggio, L. (2004a). The mirror neuron system and action recognition. *Brain Lang.* 89, 370–376.

Buccino, G., Vogt, S., Ritzl, A., Fink, G. R., Zilles, K., Freund, H. J., and Rizzolatti, G. (2004b). Neural circuits underlying imitation learning of hand actions: an event-related FMRI study. *Neuron* 42, 323–334.

Burton, M. W., Small, S. L., and Blumstein, S. E. (2000). The Role of Segmentation in phonological processing: an fMRI investigation. *J. Cogn. Neurosci.* 12, 679–690.

Callan, D. E., Jones, J. A., Callan, A. M., and Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *Neuroimage* 22, 1182–1194.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., and Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport* 14, 2213–2218.

Calvert, G. A., and Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.* 15, 57–70.

Calvert, G. A., Campbell, R., and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* 10, 649–657.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., Mcguire, P., Suckling, J., Brammer, M. J., and David, A. S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Brain Res. Cogn. Brain Res.* 12, 233–243.

Cattaneo, L., and Rizzolatti, G. (2009). The mirror neuron system. *Arch. Neurol.* 66, 557–560.

Cleveland, W. S., and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* 83, 596–610.

Cox, R. W., and Jesmanowicz, A. (1999). Real-time 3D image registration for functional MRI. *Magn. Reson. Med.* 42, 1014–1018.

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2011). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex.* PMID: 21676385. [Epub ahead of print].

D'Ausilio, A., Pulvermuller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385.

Dhanjal, N. S., Handunnetthi, L., Patel, M. C., and Wise, R. J. (2008). Perceptual systems controlling speech production. *J. Neurosci.* 28, 9969–9975.

Dick, A. S., Solodkin, A., and Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain Lang.* 114, 101–114.

Doupe, A. J., and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* 22, 567–631.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature* 384, 159–161.

Duffau, H., Capelle, L., Denvil, D., Gatignol, P., Sichez, N., Lopes, M., Sichez, J. P., and Van Effenterre, R. (2003). The role of dominant premotor cortex in language: a study using intraoperative functional mapping in awake patients. *Neuroimage* 20, 1903–1914.

Eickhoff, S. B., Heim, S., Zilles, K., and Amunts, K. (2009). A systems perspective on the effective connectivity of overt speech production. *Philos. Transact. A Math Phys. Eng. Sci.* 367, 2399–2421.

Elmer, S., Hanggi, J., Meyer, M., and Jancke, L. (2011). Differential language expertise related to white matter architecture in regions subserving sensory-motor coupling, articulation, and interhemispheric transfer. *Hum. Brain Mapp.* 32, 2064–2074.

Fabbri-Destro, M., and Rizzolatti, G. (2008). Mirror neurons and mirror systems in monkeys and humans. *Physiology (Bethesda)* 23, 171–179.

Fadiga, L., Buccino, G., Craighero, L., Fogassi, L., Gallese, V., and Pavesi, G. (1999). Corticospinal excitability is specifically modulated by motor imagery: a magnetic stimulation study. *Neuropsychologia* 37, 147–158.

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402.

Fadiga, L., Fogassi, L., Pavesi, G., and Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *J. Neurophysiol.* 73, 2608–2611.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999a). Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9, 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* 8, 272–284.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., and Dale, A. M. (2004). Automatically parcellating the human cerebral cortex. *Cereb. Cortex* 14, 11–22.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667.

Gallese, V. (2003). The manifold nature of interpersonal relations: the quest for a common mechanism. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 358, 517–528.

Gates, K. M., Molenaar, P. C., Hillary, F. G., and Slobounov, S. (2011). Extended unified SEM approach for modeling event-related fMRI data. *Neuroimage* 54, 1151–1158.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15, 870–878.

Gerbella, M., Belmalih, A., Borra, E., Rozzi, S., and Luppino, G. (2011). Cortical connections of the anterior (F5a) subdivision of the macaque ventral premotor area F5. *Brain Struct. Funct.* 216, 43–65.

Geschwind, N. (1970). The organization of language and the brain. *Science* 170, 940–944.

Gonzalez-Lima, F., and McIntosh, A. R. (1994). Neural network interactions related to auditory learning analyzed with structural equation modelling. *Hum. Brain Mapp.* 2, 23–44.

Grezes, J., Armony, J. L., Rowe, J., and Passingham, R. E. (2003). Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study. *Neuroimage* 18, 928–937.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J. Commun. Disord.* 39, 350–365.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280–301.

Hasson, U., Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron* 56, 1116–1126.

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *J. Cogn. Neurosci.* 21, 1229–1243.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402.

Holland, S. K., Plante, E., Weber Byars, A., Strawsburg, R. H., Schmithorst, V. J., and Ball, W. S. Jr. (2001). Normal fMRI brain activation patterns in children performing a verb generation task. *Neuroimage* 14, 837–843.

Iacoboni, M. (2005). Neural mechanism of imitation. *Curr. Opin. Neurobiol.* 15, 632–637.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* 3, e79. doi:10.1371/journal.pbio.0030079

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. *Front. Psychol.* 2:255. doi:10.3389/fpsyg.2011.00255

Jeannerod, M., Arbib, M. A., Rizzolatti, G., and Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends Neurosci.* 18, 314–320.

Jones, J. A., and Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport* 14, 1129–1133.

Kurata, K. (1991). Corticocortical inputs to the dorsal and ventral aspects of the premotor cortex of macaque monkeys. *Neurosci. Res.* 12, 263–280.

Lamm, C., Fischer, M. H., and Decety, J. (2007). Predicting the actions of others taps into one's own somatosensory representations – a functional MRI study. *Neuropsychologia* 45, 2480–2491.

Leff, A. P., Schofield, T. M., Stephan, K. E., Crinion, J. T., Friston, K. J., and Price, C. J. (2008). The cortical dynamics of intelligible speech. *J. Neurosci.* 28, 13209–13215.

Lotto, A. J., Hickok, G. S., and Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends Cogn. Sci. (Regul. Ed.)* 13, 110–114.

Luppino, G., Murata, A., Govoni, P., and Matelli, M. (1999). Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas AIP and VIP) and the ventral premotor cortex (areas F5 and F4). *Exp. Brain Res.* 128, 181–187.

MacLeod, A., and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *Br. J. Audiol.* 21, 131–141.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., Mcguire, P., Williams, S. C., Woll, B., and Brammer, M. J. (2000). Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport* 11, 1729–1733.

Matelli, M., Camarda, R., Glickstein, M., and Rizzolatti, G. (1986). Afferent and efferent projections of the inferior area 6 in the macaque monkey. *J. Comp. Neurol.* 251, 281–298.

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., and Mehler, J. (1993). The cortical representation of speech. *J. Cogn. Neurosci.* 5, 467–479.

McGlone, J. (1984). Speech comprehension after unilateral injection of sodium amytal. *Brain Lang.* 22, 150–157.

McIntosh, A. R. (2000). Towards a network theory of cognition. *Neural. Netw.* 13, 861–870.

McIntosh, A. R. (2004). Contexts and catalysts: a resolution of the localization and integration of function in the brain. *Neuroinformatics* 2, 175–182.

McIntosh, A. R., and Gonzalez-Lima, F. (1994). Structural equation modelling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.* 2, 2–22.

McIntosh, A. R., Grady, C. L., Ungerleider, L. G., Haxby, J. V., Rapoport, S. I., and Horwitz, B. (1994). Network analysis of cortical visual pathways mapped with PET. *J. Neurosci.* 14, 655–666.

Miller, L. M., and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 5884–5893.

Molnar-Szakacs, I., Iacoboni, M., Koski, L., and Mazziotta, J. C. (2005). Functional segregation within pars opercularis of the inferior frontal gyrus: evidence from fMRI studies of imitation and action observation. *Cereb. Cortex* 15, 986–994.

Mottonen, R., and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *J. Neurosci.* 29, 9819–9825.

Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., and Orban, G. A. (2011). Action observation circuits in the macaque monkey cortex. *J. Neurosci.* 31, 3743–3756.

Nishitani, N., and Hari, R. (2002). Viewing lip forms: cortical dynamics. *Neuron* 36, 1211–1220.

Noll, D. C., Cohen, J. D., Meyer, C. H., and Schneider, W. (1995). Spiral K-space MRI of cortical activation. *J. Magn. Reson. Imaging* 5, 49–56.

Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., and Sams, M. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage* 25, 333–338.

Ojemann, G., Ojemann, J., Lettich, E., and Berger, M. (1989). Cortical language localization in left, dominant hemisphere: an electrical stimulation mapping investigation in 117 patients. *J. Neurosurg.* 71, 316–326.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113.

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U., Sensolo, S., and Fazio, F. (2003). A functional-anatomical model for lipreading. *J. Neurophysiol.* 90, 2005–2013.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Mottonen, R., and Sams, M. (2006). Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. *Hum. Brain Mapp.* 27, 471–477.

Petrides, M., and Pandya, D. N. (1984). Projections to the frontal cortex from the posterior parietal region in the rhesus monkey. *J. Comp. Neurol.* 228, 105–116.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso Del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870.

Rizzolatti, G., and Fadiga, L. (1998). Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5). *Novartis Found. Symp.* 218, 81–95; discussion 95–103.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat. Rev. Neurosci.* 2, 661–670.

Rizzolatti, G., and Matelli, M. (2003). Two different streams form the dorsal visual system: anatomy and functions. *Exp. Brain Res.* 153, 146–157.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* 17, 1147–1153.

Rozzi, S., Ferrari, P. F., Bonini, L., Rizzolatti, G., and Fogassi, L. (2008). Functional organization of inferior parietal lobule convexity in the macaque monkey: electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. *Eur. J. Neurosci.* 28, 1569–1588.

Sato, M., Mengarelli, M., Riggio, L., Gallese, V., and Buccino, G. (2008). Task related modulation of the motor system during language processing. *Brain Lang.* 105, 83–90.

Sato, M., Tremblay, P., and Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang.* 111, 1–7.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., and Weiller, C. (2008). Ventral and dorsal pathways for language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18035–18040.

Schmahmann, J. D., Pandya, D. N., Wang, R., Dai, G., D'arceuil, H. E., De Crespigny, A. J., and Wedeen, V. J. (2007). Association fibre pathways of the brain: parallel observations from diffusion spectrum imaging and autoradiography. *Brain* 130, 630–653.

Schwartz, J. L., Basirat, A., Mènard, L., and Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguistics.* (in press).

Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neurosci. Res.* 47, 277–287.

Seltzer, B., and Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *J. Comp. Neurol.* 343, 445–463.

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89.

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2006). "Lending a helping hand to hearing: a motor theory of speech perception," in *Action To Language via the Mirror Neuron System*, ed. M. A. Arbib (Cambridge: Cambridge University Press), 250–286.

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399.

Solodkin, A., Hlustik, P., Chen, E. E., and Small, S. L. (2004). Fine modulation in network activation during motor execution and motor imagery. *Cereb. Cortex* 14, 1246–1255.

Strafella, A. P., and Paus, T. (2000). Modulation of cortical excitability during action observation: a transcranial magnetic stimulation study. *Neuroreport* 11, 2289–2292.

Sumby, W. H., and Pollack, I. (1954). Visual contribution of speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215.

Szaflarski, J. P., Schmithorst, V. J., Altaye, M., Byars, A. W., Ret, J., Plante, E., and Holland, S. K. (2006). A longitudinal functional magnetic resonance imaging study of language development in children 5 to 11 years old. *Ann. Neurol.* 59, 796–807.

Tanaka, S., and Inui, T. (2002). Cortical involvement for action imitation of hand/arm postures versus finger configurations: an fMRI study. *Neuroreport* 13, 1599–1602.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J. Cogn. Neurosci.* 17, 273–281.

Tremblay, P., Sato, M., and Small, S. L. (2011). TMS-induced modulation of action sentence priming in the ventral premotor cortex. *Neuropsychologia* 50, 319–326.

Tremblay, P., and Small, S. L. (2011). From language comprehension to action understanding and back again. *Cereb. Cortex* 21, 1166–1177.

Walsh, R. R., Small, S. L., Chen, E. E., and Solodkin, A. (2008). Network activation during bimanual movements in humans. *Neuroimage* 43, 540–553.

Watkins, K. E., Strafella, A. P., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41, 989–994.

Werker, J. F., and Tees, R. C. (1999). Influences on infant speech processing: toward a new synthesis. *Annu. Rev. Psychol.* 50, 509–535.

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702.

Wise, R. J., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K., and Warburton, E. A. (2001). Separate neural subsystems within 'Wernicke's area'. *Brain* 124, 83–95.

Wright, T. M., Pelphrey, K. A., Allison, T., Mckeown, M. J., and Mccarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex* 13, 1034–1043.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 October 2011; accepted: 05 March 2012; published online: 26 March 2012.

*Citation: Mashal N, Solodkin A, Dick AS, Chen EE and Small SL (2012) A network model of observation and imitation of speech. Front. Psychology **3**:84. doi: 10.3389/fpsyg.2012.00084*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 02 April 2012
doi: 10.3389/fpsyg.2012.00099

Gesture's neural language

***Michael Andric*[1,2] ** and *Steven L. Small*[1,2,3]**

[1] *Department of Psychology, The University of Chicago, Chicago, IL, USA*
[2] *Brain Circuits Laboratory, University of California Irvine, Irvine, CA, USA*
[3] *Department of Neurology, University of California Irvine, Irvine, CA, USA*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Thomas C. Gunter, Max-Plank-Institute for Human Cognitive and Brain Sciences, Germany
Shirley-Ann Rueschemeyer, Radboud University Nijmegen, Netherlands

****Correspondence:***
Michael Andric, Brain Circuits Laboratory, Department of Neurology, University of California Irvine, Room 3224 Biological Sciences III, Irvine, CA 92697, USA.
e-mail: andric@uchicago.edu

When people talk to each other, they often make arm and hand movements that accompany what they say. These manual movements, called "co-speech gestures," can convey meaning by way of their interaction with the oral message. Another class of manual gestures, called "emblematic gestures" or "emblems," also conveys meaning, but in contrast to co-speech gestures, they can do so directly and independent of speech. There is currently significant interest in the behavioral and biological relationships between action and language. Since co-speech gestures are actions that rely on spoken language, and emblems convey meaning to the effect that they can sometimes substitute for speech, these actions may be important, and potentially informative, examples of language–motor interactions. Researchers have recently been examining how the brain processes these actions. The current results of this work do not yet give a clear understanding of gesture processing at the neural level. For the most part, however, it seems that two complimentary sets of brain areas respond when people see gestures, reflecting their role in disambiguating meaning. These include areas thought to be important for understanding actions and areas ordinarily related to processing language. The shared and distinct responses across these two sets of areas during communication are just beginning to emerge. In this review, we talk about the ways that the brain responds when people see gestures, how these responses relate to brain activity when people process language, and how these might relate in normal, everyday communication.

Keywords: gesture, language, brain, meaning, action understanding, fMRI

INTRODUCTION

People use a variety of movements to communicate. Perhaps most familiar are the movements of the lips, mouth, tongue, and other speech articulators. However, people also perform co-speech gestures. These are arm and hand movements used to express information that accompanies and extends what is said. Behavioral research shows that co-speech gestures contribute meaning to a spoken message (Kendon, 1994; McNeill, 2005; Feyereisen, 2006; Goldin-Meadow, 2006; Hostetter, 2011). Observers integrate these gestures with ongoing speech, possibly in an automatic way (Kelly et al., 2004; Wu and Coulson, 2005). In contrast, people also use what are called emblematic gestures, or *emblems*. These are hand movements that can convey meaning directly, independent of speech (Ekman and Friesen, 1969; Goldin-Meadow, 2003). A familiar example is when someone gives a "thumbs-up" to indicate agreement or a job well done. Emblems characteristically present a conventional visual form that conveys a specific symbolic meaning, similar in effect to saying a short phrase like "Good job!" Still, both co-speech gestures and emblems are fundamentally hand actions. This is important because people encounter many types of hand actions that serve other goals and do not convey any symbolic meaning, e.g., grasping a cup. Thus, in perceiving hand actions, people routinely discern the actions' function and purpose. As this applies to understanding co-speech gestures and emblems, people must register both their manual action information and their symbolic content. It is not yet clear how the brain reconciles these manual and symbolic features. Recent research on the neurobiology involved in gesture processing implicates a variety of responses. Among these, there are responses that differentially index action and symbolic information processing. However, a characteristic response profile has yet to emerge. In what follows, we review this recent research, assess its findings in the context of the neural processing of actions and symbolic meanings, and discuss their interrelationships. We then evaluate the approaches used in prior work that has examined gesture processing. Finally, we conclude by suggesting directions for future study.

Although they are often considered uniformly, manual gestures can be classified in distinct ways. One way is by whether or not a gesture accompanies speech. Another is by the degree to which a gesture contributes meaning in its own right or in conjunction with speech. That is, manual gestures can differ in the nature of the semantic information they convey and the degree to which they rely on spoken language for their meaning. For example, *deictic* gestures provide referential information, such as when a person points to indicate "over there" and specify a location. Another class, called *beat* gestures, provide rhythm or emphasis by matching downward hand strokes with spoken intonations (McNeill, 1992). Neither deictic nor beat gestures supply semantic information in typical adult communication. In contrast, there are *iconic* and *metaphoric* gestures. These provide semantic meaning

that either complements what is said or provides information that does not otherwise come across in the verbal message (McNeill, 2005). Iconic and metaphoric gestures must be understood in the context of speech. For example, when a person moves their hand in a rolling motion, this can depict wheels turning as an iconic gesture in the context of "The wheels are turning." However, in the metaphoric use of "The meeting went on and on," the same movement can represent prolonged continuation. In other words, the speech that accompanies these gestures is key to their representational meaning.

Because gestures vary in the way they provide meaning, the relation between gestures and language is a complex, but interesting, topic. One view is that gestures and spoken language – at both psychological and biological levels of analysis – share the same communication system and are two complementary expressions of the same thought processes (McNeill, 1992). Many findings support this proposal (Cassell et al., 1999; Kelly et al., 1999; Wu and Coulson, 2005; Bernardis and Gentilucci, 2006). For example, Cassell et al. (1999) found that when people retell a narrative that was presented to them using gestures that do not match the spoken content, their retelling takes into account both the spoken and mismatched gesture information. The relation of gesture to speech is strong enough that their retelling may even include new events that resolve the conflicting speech and gestures. Moreover, another study found that when an actor pointed to an open screen door and said "The flies are out" people were much more likely to correctly understand the intended meaning (here, to close the door) when both speech and gesture were present than if only one or the other was given (Kelly et al., 1999). Thus, the way that people interpret a message is constrained when gestures and speech interact.

What are the neurobiological implications of this view that speech and gestures share a common system? There is, in fact, some neural evidence that gestures may evoke responses in brain areas that are also active when people comprehend semantic information in language. Yet, gestures are hand actions. Thus, it is also important to recognize the neural function associated with perceiving hand actions, regardless of these actions' purpose. In other words, there is a need to reconcile the neurobiology of action understanding with the neurobiology of understanding semantic information.

Prior research (described in detail below) suggests that the neural circuits involved in action understanding primarily include parts of the inferior parietal, premotor, posterior lateral temporal, and inferior frontal cortices. Interestingly, some of these brain areas, particularly in the lateral temporal and inferior frontal cortices, also respond to information conveyed in language. However, it is not known if these responses depend on the modality (e.g., language) by which this information is conveyed. Thus, this prior work leaves a number of open questions. For example, it remains unclear whether brain responses to gestures are primarily driven by the gestures' recognition as hand actions. In other words, it is uncertain whether some brain responses simply reflect sensitivity to perceiving hand actions in general, or if such responses are more tuned to the communicative information that some gestures convey. This would contrast with responses to hand actions that do not directly communicate meaning, such as grasping an object. Also, as gestures can communicate meaning, it remains to be determined if the meaning they convey is processed in a similar way as when meaning is presented in other forms, such as language. An even more basic issue is that it remains unclear whether there is a typical response profile for gestures, in general.

In the following sections, we first survey the prior research on how the brain processes manual actions, in general (*Relevant brain responses in processing gestures*). In two parts, we next review work on brain responses to gestures that communicate meaning, including emblems and co-speech gestures. We highlight areas that might respond regardless of a hand action's use in communicating meaning (*Perceiving hand actions: Inferior parietal and premotor cortex*). We then focus on brain areas thought to be important for processing meaning in language (*Perceiving meaningful hand actions: Inferior frontal and lateral temporal cortex*).

RELEVANT BRAIN RESPONSES IN PROCESSING GESTURES

People routinely perceive and understand others' hand movements. However, it is not yet clear how the brain processes such information. This is very important for understanding how gestures are recognized, since gestures are fundamentally arm and hand movements. One of the most significant findings to offer insight into a potential neural mechanism of action perception is the discovery of mirror neurons in the macaque brain. These are neurons that characteristically fire both when an animal performs a purposeful action *and* when it sees another do the same or similar act. For example, these neurons fire when the monkey sees an experimenter grasp a piece of food. They stop firing when the food is moved toward the monkey. Then, they fire again when the monkey itself grasps the food. In other words, these neurons fire in response to specific motor acts as each is perceived and performed. Mirror neurons were first found in the macaque premotor area F5 (di Pellegrino et al., 1992) and later in inferior parietal area PF (Fogassi et al., 1998). Given that area F5 receives its main parietal input from anterior PF (Geyer et al., 2000; Schmahmann et al., 2007; Petrides and Pandya, 2009), this circuit is thought to be a "parieto-frontal system that translates sensory information about a particular action into a representation of that act" (Rizzolatti et al., 1996; Fabbri-Destro and Rizzolatti, 2008). This is important because it suggests a possible a neural mechanism that would allow an "immediate, not cognitively mediated, understanding of that motor behavior" (Fabbri-Destro and Rizzolatti, 2008).

The suggestion that a "mirror mechanism" mediates action understanding in monkeys inspired attempts to try to identify a similar mechanism in humans (Rizzolatti et al., 1996, 2002; Rizzolatti and Craighero, 2004; Fabbri-Destro and Rizzolatti, 2008; Rizzolatti and Fabbri-Destro, 2008). This effort began with fMRI studies that examined brain responses when people observed grasping. Results demonstrated significant activity in premotor cortex (Buccino et al., 2001; Grezes et al., 2003; Shmuelof and Zohary, 2005, 2006), as well as parietal areas such as the intraparietal sulcus (IPS; Buccino et al., 2001, 2004; Grezes et al., 2003; Shmuelof and Zohary, 2005, 2006) and inferior parietal lobe. This also includes the supramarginal gyrus (SMG), which is thought to have some homology with monkey area PF (Perani et al., 2001; Buccino et al., 2004; Shmuelof and Zohary, 2005, 2006). For example, Buccino et al. (2001) found that when one person sees another

grasp a cup with their hand, bite an apple with their mouth, and push a pedal with their foot, not only is there parietal and premotor activity, but this activity is somatotopically organized in these areas, similar to the motor cortex homunculus (Buccino et al., 2001). Many studies (Decety et al., 1997; Grezes et al., 2003; Lui et al., 2008; Villarreal et al., 2008) find that these areas also respond when people view pantomimed actions like hammering, cutting, sawing, or using a lighter (Villarreal et al., 2008). This is particularly interesting because, with the object physically absent, it suggests that these areas respond to the action *per se* rather than to the object or to the immediate context. Furthermore, damage to these parietal and premotor areas results in damage to or loss of people's ability to produce and recognize these types of actions (Leiguarda and Marsden, 2000). However, these findings do not clarify whether such responses generalize to hand action observation, or, instead, are specific to observing actions that involve object use. In other words, would these same areas also play a functional role in understanding actions that are used to communicate meaning?

In addressing this question, several authors suggest that such a mirror mechanism might also be the basis for how the brain processes emblems and co-speech gestures (Skipper et al., 2007; Willems et al., 2007; Holle et al., 2008). However, the results needed to support this are not yet established. In particular, it is not clear whether observing a gesture systematically elicits parieto-frontal brain responses. This would be expected if a mirror mechanism based in these areas' function was integral in gesture recognition.

A further unresolved issue concerns whether these meaningful gestures elicit brain responses that are characteristically dissociable from what is found when people see hand actions that are not symbolic, such as grasping an object. These outstanding issues are considered in the following sections.

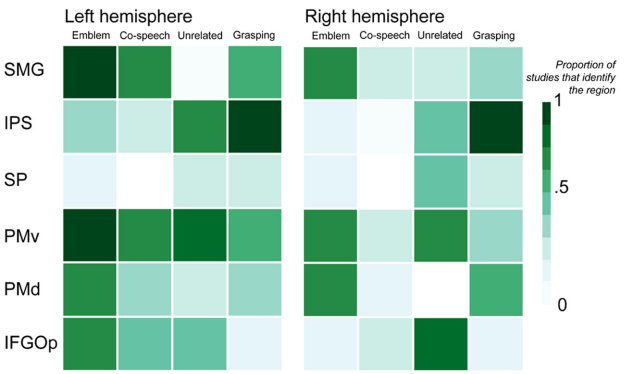


FIGURE 1 | Parietal and premotor regions relating hand action perception. Abbreviations: SMG, supramarginal gyrus; IPS, intraparietal sulcus; SP, superior parietal; PMv, ventral premotor; PMd, dorsal premotor; IFGOp, pars opercularis.

PERCEIVING HAND ACTIONS: INFERIOR PARIETAL AND PREMOTOR CORTEX

For parietal and premotor regions, their consistent (or inconsistent) reported involvement in gesture processing is illustrated in **Figure 1**. This includes results for processing *emblems*, *co-speech* gestures, hand movements that occur with speech but are *unrelated* to the spoken content, and *grasping* (see Appendix for the list of studies from which data was used to comprise the Figures). These findings most often implicate the ventral premotor cortex (PMv) and SMG as active in perceiving gestures. It is important to remember, however, that gestures can also communicate meaning. To this end, interpreting meaning (most commonly in language) is often linked to brain activity in lateral temporal and inferior frontal regions. Further below, we will address these lateral temporal and inferior frontal regions for their potential roles in gesture processing. Here, we examine parietal and premotor results.

There is evidence that parietal and premotor regions thought to be important in a putative human mirror mechanism respond not just when people view object-directed actions like grasping, but to gestures, as well. Numerous co-speech gesture (Holle et al., 2008, 2010; Dick et al., 2009; Green et al., 2009; Hubbard et al., 2009; Kircher et al., 2009; Skipper et al., 2009) and emblem (Nakamura et al., 2004; Lotze et al., 2006; Montgomery et al., 2007; Villarreal et al., 2008) studies find inferior parietal lobule activity. More precisely, the SMG and IPS are often implicated. For example, Skipper et al. (2009) found significant SMG responses when people viewed a mix of iconic, deictic, and metaphoric gestures accompanying a spoken story. In this task, the SMG also exhibited strong effective connectivity with premotor cortices (Skipper et al., 2007, 2009). Bilateral SMG activity is found when people view emblems, as well (Nakamura et al., 2004; Montgomery et al., 2007; Villarreal et al., 2008). However, the laterality of SMG effects is not consistent. For example, one study found an effect for the left SMG, but not the right SMG, when people viewed emblems (Lotze et al., 2006). The opposite was found when people saw gestures mismatched with accompanying speech (Green et al., 2009). That is, the effect was identified in the right SMG, not the left.

Both Green et al. (2009) and Willems et al. (2007) suggest that the IPS shows sensitivity when there is incongruence between gestures and speech (e.g., when a person hears "hit" and sees a "writing" gesture). However, these two studies find results in opposite hemispheres: Willems et al. (2007) identify the left IPS and Green et al. (2009) report the right IPS. Right IPS activity is also found to be stronger when people see a person make grooming or scratching movements with the hands ("adaptor movements") than when they see co-speech gestures (Holle et al., 2008). Another study, however, fails to replicate this finding (Dick et al., 2009). The right IPS is also active when people view beat gestures performed without speech (Hubbard et al., 2009). IPS activity is found in emblem studies, as well. But there is again inconsistency across reports. Whereas one study found bilateral IPS activity for processing emblems (Villarreal et al., 2008), others did not report any activity (Lotze et al., 2006; Montgomery et al., 2007). This lack of consistent IPS activity in results for co-speech and emblematic gesture processing is in contrast with results for grasping. That is, results for grasping observation consistently implicate this area. This suggests that the IPS might not play a strong role in interpreting an action's represented meaning *per se*. Rather, IPS responses

may be more tuned in processing a hand action's visuomotor properties. That is, when the focus of the presented information is the hand action, itself, (e.g., in observing grasping or beat gestures without accompanying speech) the IPS responds prominently. This would be the case also when a gesture is incongruent with accompanying speech. In this scenario, as an observer tries to reconcile divergent spoken and manual information, a more detailed processing of the hand action may be required. In contrast, when speech and gestures are congruent, processing the represented meaning, rather than the features of its expression, may be the observer's focus. In such a situation, IPS responses may not be as strong as those of other regions that are more particularly tuned toward interpreting meaning.

Premotor areas are also active when people view gestures. A number of studies report significant bilateral premotor responses to emblems (Nakamura et al., 2004; Montgomery et al., 2007; Villarreal et al., 2008). Some evidence also suggests similar premotor activity for co-speech gesture observation. For example, there is significant bilateral PMv activity when people view metaphoric gestures compared to when they view a fixation cross (Kircher et al., 2009), as well as bilateral PMd activity when they view beat gestures compared to when they watch a still body (Hubbard et al., 2009). These ventral and dorsal distinctions are also found in other studies. Specifically, whereas one study found PMd activity for emblem observation (Villarreal et al., 2008), another found activity localized to PMv, bordering the part of the inferior frontal gyrus (IFG) that also shows sensitivity to emblems (Lotze et al., 2006).

Some research suggests that premotor responses are sensitive to the semantic contribution of gesture. Willems et al. (2007) found left PMv activity to be modulated by the semantic congruency between gestures and speech: left PMv responses were stronger to gestures that were unrelated to what was said compared to when they were congruent. Another study found a similar result with gestures incongruent with a spoken homonym. However, this result implicated *both* left and right PMv cortex (Holle et al., 2008). Finally, Skipper et al. (2009) found that the BOLD signal from bilateral PMv showed a systematic response when people viewed iconic, deictic, and metaphoric gestures during audiovisual story comprehension.

Overall, these findings indicate that parietal and premotor regions are generally active when people view gestures, both as they accompany speech (co-speech gestures) and when they convey meaning without speech (emblems). Yet, parietal and premotor areas do not regularly respond in a way that indicates they are tuned specifically to whether or not the gestures convey meaning. Three primary lines of reasoning support this conclusion: (1) These areas are similarly active when people view non-symbolic actions like grasping as when they view meaningful gestures; (2) Responses in these areas do not appear to systematically distinguish between emblems and co-speech gestures, even though the former directly communicate meaning and the latter rely on speech; and (3) While some findings indicate stronger responses when a gesture does not match accompanying speech than when it does, such findings are not consistent across reports. It seems more likely that these parietal and premotor areas function more generally. That is, their responses may be evident when people view *any* purposeful hand action, rather than a specific type of hand action. In contrast, areas responsive to the meaning conveyed by these actions are more likely those thought to relate to language understanding, i.e., areas of the inferior frontal and lateral temporal cortices.

PERCEIVING MEANINGFUL HAND ACTIONS: INFERIOR FRONTAL AND LATERAL TEMPORAL CORTEX

When people see co-speech gestures with associated speech, the gestures contribute to the message's meaning and how it is interpreted (McNeill, 1992, 2005; Kendon, 1994; Goldin-Meadow, 2006). Emblems also communicate meaning. However, in contrast to co-speech gestures, they can do so directly, independent of speech. Emblems can even sometimes be used to substitute for speech (Goldin-Meadow, 1999). In fact, emblems also elicit event-related potentials comparable to those found for words (Gunter and Bach, 2004). Only recently, however, have researchers started studying how the brain processes these gestures. So far, the literature suggests both overlap and inconsistency as to which brain regions are particularly important for their processing. The variation in reported findings may be due to numerous possible sources. For example, one source may be the differing paradigms and analysis methodologies used to derive results. Another may be the extent of results that are given exposition. Also, the way that people interpret these actions may, itself, be highly variable at the neural level. Here, we highlight both the overlapping and varying findings in the gesture literature for regions in the inferior frontal and lateral temporal cortices that may be prominent in processing symbolic meaning.

Brain areas typically associated with language function also respond when people perceive gestures. This is consistent with psychological theories of gesture that propose gesture and language are two ways of expression by a single communication system (McNeill, 1992). In **Figure 2**, we illustrate the consistency of reported inferior frontal and lateral temporal region involvement in gesture processing. This again includes results for *emblems*, *co-speech* gestures, hand movements that occur with speech but are *unrelated* to the spoken content, and, for consistency with the previous figure, *grasping* (see Appendix for the studies that were used to make the figure). This figure highlights a number of areas that may be especially important in gesture processing, specifically with respect to processing the meanings they express.

The pars triangularis of the inferior frontal gyrus (IFGTr) is one area thought to be important for interpreting meaning communicated in language that might play a similar role in gesture processing. As it relates to language function, the IFGTr has been proposed to be involved in semantic retrieval and control processes when people interpret sentences and narratives (Thompson-Schill et al., 1997; Thompson-Schill, 2003). The IFGTr is consistently found to be active when people make overt semantic decisions in language tasks (Binder et al., 1997; Friederici et al., 2000; Devlin et al., 2003). A number of findings suggest this area also responds when interpreting a gesture's meaning.

The IFGTr may play a similar role in recognizing meaning from gestures as it does in verbal language. For example, bilateral IFGTr activity is stronger when people see co-speech gestures than when they process speech without gesture (Kircher et al.,

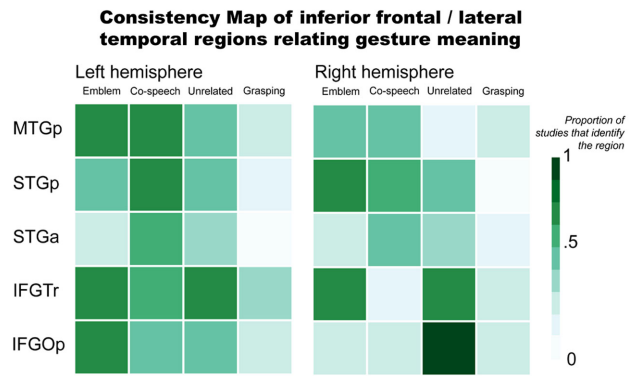


FIGURE 2 | Inferior frontal and lateral temporal regions relating gesture meaning. Abbreviations: MTGp, posterior middle temporal gyrus; STGp, posterior superior temporal gyrus; STGa, anterior superior temporal gyrus; IFGTr, pars triangularis; IFGOp, pars opercularis.

2009). Another study found greater left IFGTr activity when people observed incongruent speech and gestures than congruent (e.g., a speaker performed a "writing" gesture but said "hit"; Willems et al., 2007). However, under similar conditions in other studies – when hand movements that accompany speech are unrelated to the spoken content – activity has been found to be greater in the *right* IFGTr (Dick et al., 2009), or in both right and left IFGTr (Green et al., 2009; Straube et al., 2009). With emblems, not all studies report IFGTr activity. In those that do, though, it is found bilaterally (Lotze et al., 2006; Villarreal et al., 2008). Considered together, these results suggest that IFGTr may function similarly when people process gestures as it does when people process language. That is, IFGTr responses may be tuned to interpreting semantic information, particularly when it is necessary to unify a meaning expressed in multiple forms (e.g., via gesture and speech). When a gesture's meaning does not match speech, there is a strong IFGTr response. This may reflect added processing needed to reconcile a dominant meaning from mismatched speech and gesture. In contrast, when gesture and speech are congruent, understanding a message's meaning is more straightforward. This would rely less on regions that are particularly important for reconciling meaning from multiple representations.

Posterior to IFGTr, the pars opercularis (IFGOp) has also been found to respond when people process gestures. Anatomically positioned between IFGTr and PMv, IFGOp function has been associated with both language and motor processes. For example, this region is sensitive to audiovisual speech (Miller and D'Esposito, 2005; Hasson et al., 2007) and speech accompanied by gestures (Dick et al., 2009; Green et al., 2009; Kircher et al., 2009), as well as mouth and hand actions without any verbal communication (for review, see Binkofski and Buccino, 2004; Rizzolatti and Craighero, 2004). In other words, this area has a role in a number of language and motor functions. This includes comprehending verbal and motor information from both the mouth and the hands. Also, left frontoparietal lesions that involve the left IFG have been linked to impaired action recognition. Such impairment includes even when patients are asked to recognize an action via sounds typically associated with the action (Pazzaglia et al., 2008a,b). Put simply, the IFGOp exhibits sensitivity in response to many types of information. Such broad sensitivity suggests the IFGOp as a site where integrative processes may be important in its function.

Yet, the inferior frontal cortex functions within a broader network. During language comprehension, this area interacts with lateral temporal cortex via the extreme capsule and uncinate fasciculus fiber pathways (Schmahmann et al., 2007; Petrides and Pandya, 2009), and potentially with posterior superior temporal cortex via the superior longitudinal fasciculus (Catani et al., 2005; Glasser and Rilling, 2008). fMRI studies have described strong functional connectivity between inferior frontal and lateral temporal areas in the human brain (Homae et al., 2003; Duffau et al., 2005; Mechelli et al., 2005; Skipper et al., 2007; Saur et al., 2008; Warren et al., 2009; Xiang et al., 2010). The lateral temporal cortex also responds stronger to speech with accompanying gestures than to speech alone. In particular, the posterior superior temporal sulcus (STSp) exhibits responses to visual motion, especially when it is biologically relevant (Bonda et al., 1996; Beauchamp et al., 2002). This also applies when people perceive gestures. But according to Holle et al. (2008), responses in STSp show sensitivity beyond just perceiving biological motion. They report that left STSp is more active when people see co-speech gestures than when they see speech with adaptor movements (such as adjusting the cuff of a shirt). In a subsequent study, Holle et al. (2010) report bilateral STSp activation when people see iconic co-speech gestures compared to when people see speech, gestures alone, or to audibly degraded speech. These authors posit the left STSp as a site where "integration of iconic gestures and speech takes place." However, their effects are not replicated in other studies (e.g., Willems et al., 2007, 2009; Dick et al., 2009). For example, Dick et al. (2009) found that bilateral STSp is active both for co-speech gestures and adaptor movements. Importantly, Dick et al. (2009) did not find that activity differed between co-speech gestures and adaptor movements. That is, they did not find evidence that the STSp is responsive to the semantic content of the hand movements. This is in line with the more recognized view that the STSp is generally responsive to biological motion.

In contrast to STSp, posterior middle temporal gyrus (MTGp) and anterior superior temporal cortex (STa) responses may be tuned to interpreting meaning, including when it is conveyed in gesture. For example, bilateral MTGp activity is stronger when people see metaphoric (Kircher et al., 2009) or iconic (Green et al., 2009; Willems et al., 2009) gestures than when they see either speech or gestures alone. In response to emblems, MTGp activity has been found in each the left (Lui et al., 2008; Villarreal et al., 2008) and right (Nakamura et al., 2004) hemispheres, as well as bilaterally (Lotze et al., 2006; Xu et al., 2009). Lesion studies have also corroborated this area's importance in recognizing an action's meaning (Kalenine et al., 2010). The MTGp was considered by some authors to be part of visual association cortex (von Bonin and Bailey, 1947; Mesulam, 1985). But this region's responses to auditory stimuli are also well documented (Zatorre et al., 1992; Wise et al., 2000; Humphries et al., 2006; Hickok and Poeppel, 2007; Gagnepain et al., 2008). Many studies have also associated MTGp activity with recognizing word meaning (Binder et al., 1997; Chao et al., 1999; Gold et al., 2006). Moreover, the semantic functions

of this region might not be modality dependent (e.g., related to verbal input). That is, the MTGp may have a role in interpreting meaning at a conceptual level. This view aligns with the results of a recent meta-analysis that characterizes the MTGp as "heteromodal cortex involved in supramodal integration and conceptual retrieval" (Binder et al., 2009).

Many studies also implicate the STa in co-speech gesture processing (Skipper et al., 2007, 2009; Green et al., 2009; Straube et al., 2011). Activity in this region has been found for emblem processing, as well (Lotze et al., 2006). Changes in effective connectivity between STa and premotor cortex are found when people view gestures during story comprehension (Skipper et al., 2007). In language tasks, responses in this region have been associated with processing combinatorial meaning – usually as propositional phrases and sentences (Noppeney and Price, 2004; Humphries et al., 2006; Lau et al., 2008; Rogalsky and Hickok, 2009). A very similar function may be involved when people process gestures. After all, emblems convey propositional information that is easily translated to short spoken phrases. And co-speech gestures are typically processed in the context of sentence structures (Kircher et al., 2009) or full narratives (Skipper et al., 2007).

It appears that parts of the inferior frontal and lateral temporal cortices respond regardless of whether people perceive meaning represented verbally or manually. These areas' function suggests a shared neural basis for interpreting speech and gestures. This potentially shared basis is in line with the proposal that speech and gestures use a unified communication system (McNeill, 1992). When people must determine meaning among competing or ambiguous representations, anterior inferior frontal responses are most prevalent. In contrast, the MTGp responds strongly to represented meaning. In particular, this region appears to have a role at the level of conceptual recognition. The STa also functions in meaning recognition. Though it may be more important at the propositional level. That is, STa responses appear prominent when the expressed information involves units combined as a whole (e.g., as words are combined into phrases and sentences, or symbolic actions are associated with verbal complements).

DISTRIBUTED RESPONSES, DYNAMIC INTERACTIONS

Many reports in the gesture literature describe higher-level, complex functions (e.g., semantic integration) as localized to particular brain areas. However, the brain regularly exhibits responses that are highly distributed and specialized. These reflect the brain's dynamic functioning. Importantly, dynamic and distributed neural processes are facilitated by extensive functional connectivity and interactions. In this section, we first discuss how specialized distributed responses may apply in gesture processing, specifically in relation to motor system function. We then discuss the importance of understanding the functional relationships that facilitate cognitive processes, as these may be central in integrating and interpreting meaning from gestures and language.

The brain regularly exhibits widely distributed and diverse sets of responses. To more completely account for brain function in processing gestures, the meanings gestures convey, as well as communication in general, these distributed and diverse responses must be appreciated. Such responses may, in fact, comprise different levels of specialization that allow a functionally dynamic basis for interpretation. One view suggests that meaning, at least as it pertains to action information, is encoded via corollary processes between action and language systems (Pulvermuller, 2005; Pulvermuller et al., 2005). This view postulates that a correlation between action and action-related language leads to functional links between them. These links result in this information's encoding by distributed and interactive neural ensembles. The often cited example used to support this view is that processing effector-specific words (e.g., kick, lick, pick) involves brain activity in areas used to produce the effector-specific actions (e.g., with the leg, tongue, and mouth, respectively; Hauk et al., 2004). As reviewed above, gesture processing does, in fact, incorporate motor area responses. Whether particular types of semantic meaning conveyed by gestures is represented via distinct, distributed neural ensembles – similar to what is found for words that represent effector-specific information – is uncertain. Conceivably, motor responses in gesture processing could reflect the brain's sensitivity to represented features, beyond a gesture's visuomotor properties. In other words, gestures that symbolically represent motor information (e.g., a gesture used to represent a specific body part, such as the leg) could also rely on a somatotopic encoding analogous to that found for words that represent body parts. The way that the brain would achieve this degree of specialization is uncertain. It is increasingly clear though that responses to gestures are, indeed, diverse and distributed among distinct regions (**Figures 1** and **2**). Yet, this view that action information is encoded via corollary processes between action and language systems does not account for processing meaning that does not involve action (e.g., "The capitol is Sacramento"). Thus, while distributed encoding in the motor system may play a part in processing information that relates actions to the effectors used to perform actions, accounting for how the brain interacts with meaning more generally requires a broader basis.

Understanding how the brain interprets gestures and the information they convey requires appreciating the way that the brain represents information and implements higher-level functions. This requires characterizing not just the function of distinct locations that may show tuning to particular features, but also the dynamic interactions and aggregate function of distributed responses (McIntosh, 2000). Certain brain areas (e.g., in sensory and motor cortices) may be specialized to respond to particular kinds of information. But higher-level processes, such as memory and language (and by extension, interpreting meaning), require understanding the way that the brain relates and integrates information. Most of the previously discussed studies, particularly those focused on co-speech gestures, have aimed to characterize the neural integration of gesture and language processing. Many localize this process with results for sets of individual regions. Some of the implicated regions include the IFG (Willems et al., 2007; Straube et al., 2009), temporo-occipital junction (Green et al., 2009), and STSp (Holle et al., 2008, 2010). However, to characterize complex, integrative functions by one-to-one alliances with individual regions, without also acknowledging those neural mechanisms that might enable relationships among particular regions, loses sight of the brain's dynamic and interconnected nature. For example, the same brain areas may exhibit activity in different tasks or in response to similar information from different

mediums (e.g., each symbolic gestures and spoken language; Xu et al., 2009). Similarly, the brain can exhibit distributed function that is evoked by presentation from the same medium (e.g., co-speech gestures). Thus, whereas a particular brain area may be similarly active across what seem to be different cognitive tasks, what "distinguishes [these] tasks is the pattern of spatiotemporal activity and interactivity more than the participation of any particular region" (McIntosh, 2000). Some previous gesture work has examined the functional relationships among anatomically diverse areas' responses (e.g., Skipper et al., 2009; Willems et al., 2009; Xu et al., 2009). However, such analyses are still the exception in the gesture literature. A more global perspective that recognizes specialized responses interact across the whole brain to implement cognitive processes is needed. That is, whereas one neural system might be particularly tuned to process gestures, another might be better tuned to process verbal discourse. Importantly, while such systems may organize with varying degrees of specialization, it is their dense interconnectivity that enables dynamic neural processing in context. This may be especially important for understanding the way that the brain functions in the perceptually rich and complex scenarios that comprise typical experience. Thus, to understand the way the brain implements complex processes, such as integrating and interpreting meaning from gestures and language, "considering activity of the entire brain rather than individual regions" (McIntosh, 2000) is vital.

RELEVANCE FOR REAL WORLD INTERACTIONS (BEYOND THE EXPERIMENT)

To understand a gesture, an observer must relate multiple pieces of information. For example, emblem comprehension involves visually perceiving the gesture, as well as processing its meaning. Similarly, interpreting co-speech gestures requires visual perception of the gesture. But, in contrast with emblems, co-speech gesture processing involves associating the action with accompanying auditory verbal information. Importantly, speech and gesture information do not combine in an additive way. Rather, these sources *interactively* contribute meaning, as people integrate them into a unified message (Kelly et al., 1999; Bernardis and Gentilucci, 2006; Gentilucci et al., 2006). There is also pragmatic information that is part of the natural context in which these actions are typically experienced. This pragmatic context can also influence a gesture's interpretation (Kelly et al., 1999, 2007). Another factor that can impact interpretation is the observer's intent. For example, brain responses to the same gestures can differ depending on whether an observer's goal is to recognize the hand as meaningful or, categorically, as simply a hand (Nakamura et al., 2004). Therefore, to comprehensively appreciate the way that the brain processes gestures – particularly, the meaning they express – these diverse information sources should be accounted for in ways that recognize contextual influences.

However, most fMRI studies of gesture processing present participants with stimuli that have little or no resemblance to anything they would encounter outside of the experiment. Of course, researchers do this with the intention of isolating brain responses to a specific feature or function of interest by controlling for all other factors. Some examples in prior gesture studies include having the person performing the gestures cloaked in all black (Holle et al., 2008), allowing only the actor's hand to be visible through a screen (Montgomery et al., 2007), and putting a large circle that changes colors on the actor's chest (Kircher et al., 2009). Such unusual visual information could pose a number of problems. Most concerning is that it could distract attention from the visual information that is relevant and of interest (e.g., the gestures). The inverse is also possible, however. That is, irregular visual materials might artificially enhance attention toward the gestures. In either case, such materials do not generalize to people's typical experience.

Beyond the materials' visual aspects, many experiments have also used conditions that are explicitly removed from familiar experience. For example, a common approach has been to compare responses to co-speech gestures with responses to speech and gestures that do not match. The idea here is that the difference between these conditions would reveal brain areas involved in "integration" (itself often only loosely or not at all defined). The reasoning is that in the condition where gesture and speech are mismatched the brain is presumed to respond to each as dissociated signals. But, when gesture and speech are congruent, there is recognition of a unified representation or message. However, meaningless hand actions evoke a categorically different brain response than meaningful ones (Decety et al., 1997). Thus, interpreting these findings can be difficult. Such an approach also brings to light an additional potential limitation: Many results are determined by simply subtracting responses collected in one condition from those in another condition. In other words, results are often achieved by subtracting responses generated under exposure to one input (e.g., speech) from responses to a combination of inputs (e.g., speech and gesture). The difference in activity for the contrast or condition of interest is then typically described as the effect. Not only does such an approach make it difficult to characterize the interaction between speech and gesture that gives a co-speech gesture meaning, but it also assumes each the brain and fMRI signals are linear (which they are not; Logothetis et al., 2001). Thus, results derived under such conditions can be hard to interpret, particularly as to the degree to which they inform gesture processing. They may also yield results that are hard to replicate, even when a study explicitly tries to do so (Dick et al., 2009).

Another issue is that many researchers require their participants to do motor tasks (such as pushing buttons to record behaviors) that are accessory to the function of interest during fMRI data collection. A number of prior gesture studies have used these tasks (e.g., Green et al., 2009; Kircher et al., 2009). However, having participants engage in motor behaviors while in the scanner could potentiate responses in a confounding way. In other words, motor responses could then be due to the motor behavior in the accessory task, as well as interfere with potential motor responses relevant to processing the gestures (the "motor output problem"; Small and Nusbaum, 2004). This can be especially problematic when motor areas are of primary interest. Thus, these accessory tasks can generate brain responses that are hard to disentangle from those that the researchers intended to examine.

To avoid many of these potential limitations, more naturalistic conditions need to be considered in studying gesture and language function. One immediate concern may be that evaluating data collected under contextualized, more naturalistic exposures can pose

a challenge for commonly used fMRI analysis measures. The most typical approach for analyzing fMRI data uses the general linear model. This requires an *a priori* specified hemodynamic response model against which the collected responses can be regressed. Also, particularly for event-related designs, an optimized stimulus event sequence is needed to avoid co-linearity effects that may mask signal of interest from co-varying noise-related artifacts. With naturalistic, continuously unfolding stimuli, meeting such requirements is not always possible. Fortunately, many previous authors have demonstrated approaches that achieve systematic, informative results from data collected under more naturalistic conditions (e.g., Zacks et al., 2001; Bartels and Zeki, 2004; Hasson et al., 2004; Mathiak and Weber, 2006; Malinen et al., 2007; Spiers and Maguire, 2007; Yarkoni et al., 2008; Skipper et al., 2009; Stephens et al., 2010). The intersubject synchronization approach used to analyze data collected while people watched segments of "The Good, The Bad, and The Ugly" is probably the most well-known example (Hasson et al., 2004). However, others researchers have successfully derived informative fMRI results from data collected as people comprehended naturalistic audiovisual stories (Wilson et al., 2008; Skipper et al., 2009), read narratives (Yarkoni et al., 2008), watched videos that presented everyday events such as doing the dishes (Zacks et al., 2001), and had verbal communication in the scanner (Stephens et al., 2010). These achievements are important because they demonstrate ways to gainfully use context rather than unnaturally remove it. Considering the interactive, integrative, and contextual nature of gesture and language processing – particularly in typical experience – it is essential to consider such approaches as the study of gesture and language moves forward.

SUMMARY AND FUTURE PERSPECTIVES

SUMMARY

Current findings indicate that two types of brain areas are implicated when people process gestures, particularly emblems and co-speech gestures. One set of areas comprises parietal and premotor regions that are important in processing hand actions. These areas are sensitive to both gestures and actions that are not directly symbolic, such as grasping an object. Thus, it is likely that the function of these parietal and premotor regions primarily involves perceiving hand actions, rather than interpreting their meaning. In contrast, the other set of areas includes inferior frontal and lateral temporal regions. These regions are classically associated with language processing. They may function in a similar capacity to process symbolic meaning conveyed with gestures. While the current data present a general consensus for these areas' roles, the way that the brain reconciles manual and symbolic information it is not yet clear.

The lack of reconciliation between the neurobiology of action understanding and that of understanding symbolic information yields at least two prominent points concerning gesture research. First, a characteristic response profile for gesture processing remains unspecified. That is, among results for these two sets of areas, there is a strong degree of variability. This variability clouds whether certain regions' responses are central in gesture processing. It also obscures whether there are particular sets of responses that implement the integrative and interpretive mechanisms needed to comprehend gestures and language. Second, there is currently minimal exposition at the neural systems level, particularly that relates the brain's anatomical and functional interconnections. The variable and widely distributed responses found in previous gesture studies suggest a broader neural perspective is needed. In other words, function throughout the brain and its interconnectivity must be considered. In the final section, we discuss important issues for moving forward in the study of gesture and language processing and then relate them to some outstanding topics.

CONTEXT IS PERVASIVE

To move forward in understanding the way that the brain processes gestures and language, the fundamental importance of context must be recognized. This pertains both to experimental design and as a principle of brain function ("neural context"; McIntosh, 2000). Here, we will first briefly summarize the importance of considering context as it relates to experimental approaches. We then follow with a discussion of context as it relates to understanding interactive and dynamic function across the whole brain.

Concerning the role of context in experimental design, one of the primary hurdles in using contextualized, naturalistic materials is that they do not typically satisfy the *a priori* requirements of many commonly used imaging analysis methods (as discussed above). However, systematically evaluating fMRI data collected under contextualized, continuous exposures is achievable. Prior imaging work includes numerous informative results derived from fMRI data collected under more naturalistic conditions (Zacks et al., 2001; Bartels and Zeki, 2004; Hasson et al., 2004; Mathiak and Weber, 2006; Malinen et al., 2007; Spiers and Maguire, 2007; Yarkoni et al., 2008; Skipper et al., 2009; Stephens et al., 2010). Gesture and language researchers should increasingly consider such methods. Applying these methods could supplement subtractive approaches (both in designing experimental conditions and their analysis) that might mischaracterize the way the brain operates. In particular, these methods might provide better insight into integrative mechanisms in gesture and language processing, as they are implemented in typical experience.

Concerning the role of context as it relates to brain function, it is important to maintain perspective of the entire brain's function. Gesture and language processing exemplifies this need, as they incorporate responses that are not only diverse and distributed across the brain but are also interactive and interconnected. In other words, recognizing that the brain is a complex system in the formal (mathematical) sense will benefit efforts to understand gesture and language function. Accordingly, this will necessitate investigations to focus on distributed neural systems, rather than just on localizing complex processes to individual regions. Examining the neural relationships between regions could more comprehensively characterize gesture and language processing. In any case, it would promote research that better investigates those neural properties that facilitate higher-level functions. For example, the distributed brain networks involved in functions such as language, memory, and attention comprise multiple pathways (Mesulam, 1990) with semi-redundant and reciprocal connectivity (Tononi and Sporns, 2003; Friston, 2005). These connections may involve regions with varying degrees

of specialization to particular information types (e.g., gesture and/or language stimuli). Also, a particular region's specialization may be determined, in part, by its connectivity (McIntosh, 2000). Put simply, more than one region may be engaged in a particular function, and more than one function may engage a particular region. It is important then to consider that "specialization is only meaningful in the context of functional integration and vice versa" (Friston, 2005). Therefore, to appreciate how the brain implements complex operations (e.g., information integration, interpretation), insight into neural function at multiple levels of representation is needed. This implicates not just the level of regional specialization (the system's elements) but also the anatomical and functional relationships that facilitate these elements' interactive and distributed processes (their connections).

FUNCTIONALLY INTERACTIVE AND DISTRIBUTED OPERATIONS

Recognizing the importance of context – both as it relates to experimental design and as a basic principle of brain function – will allow future studies to better examine how the brain operates through interactive and distributed function. This will encourage researchers to ask questions about the brain that incorporate contextual factors, rather than artificially eliminate them. This is important because people function in a world that requires continual interaction with abundant, changing, and diverse information sources. A greater degree of resemblance and relevance to the real world can and should be incorporated into future experimental designs. Below, we consider a few outstanding issues in the study of gesture and language processing for which these approaches may be especially useful.

One issue is whether the neural mechanisms involved in perceiving gestures distinguish among the diverse semantic meanings they can represent. For example, co-speech gestures can be used to represent different types of information such as physical objects, body parts, or abstracts ideas. It is unclear if the brain exhibits specialized responses that are particular to these meanings. Whether or not such responses might be distributed in distinct regions is also uncertain. Additionally, it is unclear how distributed responses to gestures would interact with other neural systems to incorporate contextual factors.

Another issue is to what extent pragmatics and situational factors influence how brain systems organize in processing gesture and language information. Pragmatic knowledge does, in fact, play a role in gesture comprehension (Kelly et al., 1999, 2007). It is uncertain, however, to what degree certain systems, such as the putative action understanding circuit relating parietal and premotor responses, would maintain a functional role in perceiving gestures under varying situational influences. For example, a person might perceive a pointing gesture by someone yelling "Over there!" while trying to escape a burning building. This is quite different, and would probably involve different neural systems, than perceiving the same gesture and language conveyed to indicate where the TV remote is located. Similarly, the way that neural mechanisms coordinate as a function of the immediate verbal context is also uncertain. Recall the example provided in the Introduction of this paper: A person moving their hand in a rolling motion can represent one thing accompanied by "The wheels are turning" but another when accompanied by "The meeting went on and on." To process such information, neural mechanisms that enable functional interaction among responses to the gesture and accompanying verbal content, and that dynamically implement interpreting their meaning in context, would need to be incorporated. Such interactive neural mechanisms are currently unclear and deserve further investigation. Thus, addressing these issues would further inform the neural basis of gesture processing, as well as how the brain might encode and interpret meaning.

In conclusion, the current gesture data implicate a number of brain areas that to differing extents index action and symbolic information processing. However, the neural relationships that provide a dynamic and interactive basis for comprehension need to be accounted for, as well. This will allow a more complete look at the way that the brain processes gestures and their meanings. As gesture and language research moves forward, a vital factor that needs further consideration is the role of context. The importance of context applies both to the settings in which people process gesture and language information, as well as to understanding in what way distributed and interconnected responses throughout the brain facilitate this information's interactive comprehension.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Deafness and Other Communication Disorders NIH R01-DC03378.

REFERENCES

Bartels, A., and Zeki, S. (2004). Functional brain mapping during free viewing of natural scenes. *Hum. Brain Mapp.* 21, 75–85.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., and Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron* 34, 149–159.

Bernardis, P., and Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia* 44, 178–190.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., and Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *J. Neurosci.* 17, 353–362.

Binkofski, F., and Buccino, G. (2004). Motor functions of the Broca's region. *Brain Lang.* 89, 362–369.

Bonda, E., Petrides, M., Ostry, D., and Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *J. Neurosci.* 16, 3737–3744.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., and Freund, H. J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *Eur. J. Neurosci.* 13, 400–404.

Buccino, G., Binkofski, F., and Riggio, L. (2004). The mirror neuron system and action recognition. *Brain Lang.* 89, 370–376.

Cassell, J., McNeill, D., and McCullough, K.-E. (1999). Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics Cogn.* 7, 1–34.

Catani, M., Jones, D. K., and Ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Ann. Neurol.* 57, 8–16.

Chao, L. L., Haxby, J. V., and Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat. Neurosci.* 2, 913–919.

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., and Fazio, F. (1997). Brain activity during observation of actions – influence of action content and subject's strategy. *Brain* 120, 1763–1777.

Devlin, J. T., Matthews, P. M., and Rushworth, M. F. S. (2003). Semantic processing in the left inferior prefrontal cortex: a combined functional magnetic resonance imaging and transcranial magnetic stimulation study. *J. Cogn. Neurosci.* 15, 71–84.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events – a neurophysiological study. *Exp. Brain Res.* 91, 176–180.

Dick, A. S., Goldin-Meadow, S., Hasson, U., Skipper, J. I., and Small, S. L. (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Hum. Brain Mapp.* 30, 3509–3526.

Duffau, H., Gatignol, P., Mandonnet, E., Peruzzi, P., Tzourio-Mazoyer, N., and Capelle, L. (2005). New insights into the anatomo-functional connectivity of the semantic system: a study using cortico-subcortical electrostimulations. *Brain* 128, 797–810.

Ekman, P., and Friesen, W. V. (1969). The repertoire of nonverbal communication: categories, origins, usage, and coding. *Semiotica* 1, 49–98.

Fabbri-Destro, M., and Rizzolatti, G. (2008). Mirror neurons and mirror systems in monkeys and humans. *Physiology (Bethesda)* 23, 171–179.

Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: their meaning matters. *Eur. J. Cogn. Psychol.* 18, 185–205.

Fogassi, L., Gallese, V., Fadiga, L., and Rizzolatti, G. (1998). Neurons responding to the sight of goal directed hand/arm actions in the parietal area PF (7b) of the macaque monkey. *Soc. Neurosci. Abstr.* 24, 257.5.

Friederici, A. D., Opitz, B., and Cramon, D. Y. V. (2000). Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cereb. Cortex* 10, 698–705.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836.

Gagnepain, P., Chetelat, G., Landeau, B., Dayan, J., Eustache, F., and Lebreton, K. (2008). Spoken word memory traces within the human auditory cortex revealed by repetition priming and functional magnetic resonance imaging. *J. Neurosci.* 28, 5281–5289.

Gentilucci, M., Bernardis, P., Crisi, G., and Dalla Volta, R. (2006). Repetitive transcranial magnetic stimulation of Broca's area affects verbal responses to gesture observation. *J. Cogn. Neurosci.* 18, 1059–1074.

Geyer, S., Matelli, M., Luppino, G., and Zilles, K. (2000). Functional neuroanatomy of the primate isocortical motor system. *Anat. Embryol.* 202, 443–474.

Glasser, M. F., and Rilling, J. K. (2008). DTI tractography of the human brain's language pathways. *Cereb. Cortex* 18, 2471–2482.

Gold, B. T., Balota, D. A., Jones, S. J., Powell, D. K., Smith, C. D., and Andersen, A. H. (2006). Dissociation of automatic and strategic lexical-semantics: functional magnetic resonance imaging evidence for differing roles of multiple frontotemporal regions. *J. Neurosci.* 26, 6523–6532.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends Cogn. Sci. (Regul. Ed.)* 3, 419–429.

Goldin-Meadow, S. (2003). *Hearing Gesture: How our Hands Help us Think.* Cambridge, MA: Belknap Press of Harvard University Press.

Goldin-Meadow, S. (2006). Talking and thinking with our hands. *Curr. Dir. Psychol. Sci.* 15, 34–39.

Grafton, S. T., Arbib, M. A., Fadiga, L., and Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. *Exp. Brain Res.* 112, 103–111.

Green, A., Straube, B., Weis, S., Jansen, A., Willmes, K., Konrad, K., and Kircher, T. (2009). Neural integration of iconic and unrelated coverbal gestures: a functional MRI study. *Hum. Brain Mapp.* 30, 3309–3324.

Grezes, J., Armony, J. L., Rowe, J., and Passingham, R. E. (2003). Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study. *Neuroimage* 18, 928–937.

Gunter, T. C., and Bach, P. (2004). Communicating hands: ERPs elicited by meaningful symbolic hand postures. *Neurosci. Lett.* 372, 52–56.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634–1640.

Hasson, U., Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron* 56, 1116–1126.

Hauk, O., Johnsrude, I., and Pulvermuller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41, 301–307.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402.

Holle, H., Gunter, T. C., Rüschemeyer, S. A., Hennenlotter, A., and Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage* 39, 2010–2024.

Holle, H., Obleser, J., Rueschemeyer, S.-A., and Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *Neuroimage* 49, 875–884.

Homae, F., Yahata, N., and Sakai, K. L. (2003). Selective enhancement of functional connectivity in the left prefrontal cortex during sentence processing. *Neuroimage* 20, 578–586.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychol. Bull.* 137, 297–315.

Hubbard, A. L., Wilson, S. M., Callan, D. E., and Dapretto, M. (2009). Giving speech a hand: gesture modulates activity in auditory cortex during speech perception. *Hum. Brain Mapp.* 30, 1028–1037.

Humphries, C., Binder, J. R., Medler, D. A., and Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J. Cogn. Neurosci.* 18, 665–679.

Kalenine, S., Buxbaum, L. J., and Coslett, H. B. (2010). Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke. *Brain* 133, 3269–3280.

Kelly, S. D., Barr, D. J., Church, R. B., and Lynch, K. (1999). Offering a hand to pragmatic understanding: the role of speech and gesture in comprehension and memory. *J. Mem. Lang.* 40, 577–592.

Kelly, S. D., Kravitz, C., and Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain Lang.* 89, 253–260.

Kelly, S. D., Ward, S., Creigh, P., and Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain Lang.* 101, 222–233.

Kendon, A. (1994). Do gestures communicate? A review. *Res. Lang. Soc. Interact.* 27, 175–200.

Kircher, T., Straube, B., Leube, D., Weis, S., Sachs, O., Willmes, K., Konrad, K., and Green, A. (2009). Neural interaction of speech and gesture: differential activations of metaphoric co-verbal gestures. *Neuropsychologia* 47, 169–179.

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* 9, 920–933.

Leiguarda, R. C., and Marsden, C. D. (2000). Limb apraxias: higher-order disorders of sensorimotor integration. *Brain* 123(Pt 5), 860–879.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150–157.

Lotze, M., Heymans, U., Birbaumer, N., Veit, R., Erb, M., Flor, H., and Halsband, U. (2006). Differential cerebral activation during observation of expressive gestures and motor acts. *Neuropsychologia* 44, 1787–1795.

Lui, F., Buccino, G., Duzzi, D., Benuzzi, F., Crisi, G., Baraldi, P., Nichelli, P., Porro, C. A., and Rizzolatti, G. (2008). Neural substrates for observing and imagining non-object-directed actions. *Soc. Neurosci.* 3, 261–275.

Malinen, S., Hlushchuk, Y., and Hari, R. (2007). Towards natural stimulation in fMRI – issues of data analysis. *Neuroimage* 35, 131–139.

Manthey, S., Schubotz, R. I., and Von Cramon, D. Y. (2003). Premotor cortex in observing erroneous action: an fMRI study. *Brain Res. Cogn. Brain Res.* 15, 296–307.

Mathiak, K., and Weber, R. (2006). Toward brain correlates of natural behavior: fMRI during violent video games. *Hum. Brain Mapp.* 27, 948–956.

McIntosh, A. R. (2000). Towards a network theory of cognition. *Neural Netw.* 13, 861–870.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought.* Chicago: University of Chicago Press.

Mechelli, A., Crinion, J. T., Long, S., Friston, K. J., Lambon Ralph, M. A., Patterson, K., McClelland, J. L., and Price, C. J. (2005). Dissociating reading processes on the basis of neuronal interactions. *J. Cogn. Neurosci.* 17, 1753–1765.

Mesulam, M. M. (1985). "Patterns in behavioral neuroanatomy: association areas, the limbic system, and hemispheric specialization," in

Principles of Behavioral Neurology, ed. M. M. Mesulam (Philadelphia: F. A. Davis), 1–70.

Mesulam, M.-M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Ann. Neurol.* 28, 597–613.

Miller, L. M., and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 5884–5893.

Montgomery, K. J., Isenberg, N., and Haxby, J. V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Soc. Cogn. Affect. Neurosci.* 2, 114–122.

Nakamura, A., Maess, B., Knosche, T. R., Gunter, T. C., Bach, P., and Friederici, A. D. (2004). Cooperation of different neuronal systems during hand sign recognition. *Neuroimage* 23, 25–34.

Noppeney, U., and Price, C. J. (2004). An FMRI study of syntactic adaptation. *J. Cogn. Neurosci.* 16, 702–713.

Pazzaglia, M., Pizzamiglio, L., Pes, E., and Aglioti, S. M. (2008a). The sound of actions in apraxia. *Curr. Biol.* 18, 1766–1772.

Pazzaglia, M., Smania, N., Corato, E., and Aglioti, S. M. (2008b). Neural underpinnings of gesture discrimination in patients with limb apraxia. *J. Neurosci.* 28, 3030–3041.

Perani, D., Fazio, F., Borghese, N. A., Tettamanti, M., Ferrari, S., Decety, J., and Gilardi, M. C. (2001). Different brain correlates for watching real and virtual hand actions. *Neuroimage* 14, 749–758.

Petrides, M., and Pandya, D. N. (2009). Distinct parietal and temporal pathways to the homologues of Broca's area in the monkey. *PLoS Biol.* 7, e1000170. doi:10.1371/journal.pbio.1000170

Pierno, A. C., Becchio, C., Wall, M. B., Smith, A. T., Turella, L., and Castiello, U. (2006). When gaze turns into grasp. *J. Cogn. Neurosci.* 18, 2130–2137.

Pulvermuller, F. (2005). Brain mechanisms linking language and action. *Nat. Rev. Neurosci.* 6, 576–582.

Pulvermuller, F., Hauk, O., Nikulin, V. V., and Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *Eur. J. Neurosci.* 21, 793–797.

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192.

Rizzolatti, G., and Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Curr. Opin. Neurobiol.* 18, 179–184.

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Res. Cogn. Brain Res.* 3, 131–141.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2002). Motor and cognitive functions of the ventral premotor cortex. *Curr. Opin. Neurobiol.* 12, 149–154.

Rogalsky, C., and Hickok, G. (2009). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cereb. Cortex* 19, 786–796.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., and Weiller, C. (2008). Ventral and dorsal pathways for language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18035–18040.

Schmahmann, J. D., Pandya, D. N., Wang, R., Dai, G., D'Arceuil, H. E., de Crespigny, A. J., and Wedeen, V. J. (2007). Association fibre pathways of the brain: parallel observations from diffusion spectrum imaging and autoradiography. *Brain* 130, 630–653.

Shmuelof, L., and Zohary, E. (2005). Dissociation between ventral and dorsal fMRI activation during object and action recognition. *Neuron* 47, 457–470.

Shmuelof, L., and Zohary, E. (2006). A mirror representation of others' actions in the human anterior parietal cortex. *J. Neurosci.* 26, 9736–9742.

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., and Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain Lang.* 101, 260–277.

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., and Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Curr. Biol.* 19, 1–7.

Small, S. L., and Nusbaum, H. C. (2004). On the neurobiological investigation of language understanding in context. *Brain Lang.* 89, 300–311.

Spiers, H. J., and Maguire, E. A. (2007). Decoding human brain activity during real-world experiences. *Trends Cogn. Sci. (Regul. Ed.)* 11, 356–365.

Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14425–14430.

Straube, B., Green, A., Bromberger, B., and Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: common and unique integration processes. *Hum. Brain Mapp.* 32, 520–533.

Straube, B., Green, A., Weis, S., Chatterjee, A., and Kircher, T. (2009). Memory effects of speech and gesture binding: cortical and hippocampal activation in relation to subsequent memory performance. *J. Cogn. Neurosci.* 21, 821–836.

Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: inferring "how" from "where". *Neuropsychologia* 41, 280–292.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., and Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proc. Natl. Acad. Sci. U.S.A.* 94, 14792–14797.

Tononi, G., and Sporns, O. (2003). Measuring information integration. *BMC Neurosci.* 4, 31. doi:10.1186/1471-2202-4-31

Tremblay, P., and Small, S. L. (2011). From language comprehension to action understanding and back again. *Cereb. Cortex* 21, 1166–1177.

Villarreal, M., Fridman, E. A., Amengual, A., Falasco, G., Gerscovich, E. R., Ulloa, E. R., and Leiguarda, R. C. (2008). The neural substrate of gesture recognition. *Neuropsychologia* 46, 2371–2382.

von Bonin, G., and Bailey, P. (1947). *The Neocortex of Macaca Mulatta*. Urbana: University of Illinois Press.

Warren, J. E., Crinion, J. T., Lambon Ralph, M. A., and Wise, R. J. (2009). Anterior temporal lobe connectivity correlates with functional outcome after aphasic stroke. *Brain* 132, 3428–3442.

Willems, R. M., Ozyurek, A., and Hagoort, P. (2007). When language meets action: the neural integration of gesture and speech. *Cereb. Cortex* 17, 2322–2333.

Willems, R. M., Ozyurek, A., and Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *Neuroimage* 47, 1992–2004.

Wilson, S. M., Molnar-Szakacs, I., and Iacoboni, M. (2008). Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. *Cereb. Cortex* 18, 230–242.

Wise, R. J., Howard, D., Mummery, C. J., Fletcher, P., Leff, A., Buchel, C., and Scott, S. K. (2000). Noun imageability and the temporal lobes. *Neuropsychologia* 38, 985–994.

Wu, Y. C., and Coulson, S. (2005). Meaningful gestures: electrophysiological indices of iconic gesture comprehension. *Psychophysiology* 42, 654–667.

Xiang, H. D., Fonteijn, H. M., Norris, D. G., and Hagoort, P. (2010). Topographical functional connectivity pattern in the Perisylvian language networks. *Cereb. Cortex* 20, 549–560.

Xu, J., Gannon, P. J., Emmorey, K., Smith, J. F., and Braun, A. (2009). Symbolic gestures and spoken language are processed by a common neural system. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20664–20669.

Yarkoni, T., Speer, N. K., and Zacks, J. M. (2008). Neural substrates of narrative comprehension and memory. *Neuroimage* 41, 1408–1425.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., and Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nat. Neurosci.* 4, 651–655.

Zatorre, R. J., Evans, A. C., Meyer, E., and Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science* 256, 846–849.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 September 2011; accepted: 16 March 2012; published online: 02 April 2012.

Citation: Andric M and Small SL (2012) Gesture's neural language. Front. Psychology ***3****:99. doi: 10.3389/fpsyg.2012.00099*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

APPENDIX

STUDIES USED IN FIGURES 1 AND 2

Emblematic gesture

1. Lotze et al. (2006)
2. Lui et al. (2008)
3. Montgomery et al. (2007)
4. Nakamura et al. (2004)
5. Villarreal et al. (2008)

Co-speech gesture

1. Dick et al. (2009)
2. Green et al. (2009)
3. Holle et al. (2008)
4. Holle et al. (2010)
5. Hubbard et al. (2009)
6. Kircher et al. (2009)
7. Skipper et al. (2007)
8. Skipper et al. (2009)
9. Straube et al. (2009)
10. Straube et al. (2011)
11. Willems et al. (2007)
12. Willems et al. (2009)

Unrelated co-speech gesture

1. Dick et al. (2009)
2. Green et al. (2009)
3. Holle et al. (2008)
4. Hubbard et al. (2009)
5. Straube et al. (2009)
6. Willems et al. (2007)
7. Willems et al. (2009)

Grasping

1. Buccino et al. (2001)
2. Buccino et al. (2004)
3. Grezes et al. (2003)
4. Grafton et al. (1996)
5. Shmuelof and Zohary (2005)
6. Shmuelof and Zohary (2006)
7. Manthey et al. (2003)
8. Perani et al. (2001)
9. Pierno et al. (2006)
10. Tremblay and Small (2011)

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 02 April 2012
doi: 10.3389/fpsyg.2012.00081

Cognitive and electrophysiological correlates of the bilingual Stroop effect

Lavelda J. Naylor[1], Emily M. Stanley[2] and Nicole Y. Y. Wicha[1,3,4]*

[1] *Department of Biology, University of Texas at San Antonio, San Antonio, TX, USA*
[2] *Department of Psychology, University of Delaware, Newark, DE, USA*
[3] *Neurosciences Institute, University of Texas at San Antonio, San Antonio, TX, USA*
[4] *Research Imaging Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Lucy Jane MacGregor, Medical Research Council, UK
Kira Bailey, Iowa State University, USA

****Correspondence:***
Nicole Y. Y. Wicha, Department of Biology, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA.
e-mail: nicole.wicha@utsa.edu

The color word Stroop effect in bilinguals is commonly half the magnitude when the written and naming languages are different (between) than when they are the same (within). This between-within language Stroop difference (BWLS) is likened to a response set effect, with greater response conflict for response relevant than irrelevant words. The nature of the BWLS was examined using a bilingual Stroop task. In a given block (Experiment 1), color congruent and incongruent words appeared in the naming language or not (single), or randomly in both languages (mixed). The BWLS effect was present for both balanced and unbalanced bilinguals, but only partially supported a response set explanation. As expected, color incongruent trials during single language blocks, lead to slower response times within than between languages. However, color congruent trials during mixed language blocks led to slower times between than within languages, indicating that response-irrelevant stimuli interfered with processing. In Experiment 2, to investigate the neural timing of the BWLS effect, event related potentials were recorded while balanced bilinguals named silently within and between languages. Replicating monolingual findings, an N450 effect was observed with larger negative amplitude for color incongruent than congruent trials (350–550 ms post-stimulus onset). This effect was equivalent within and between languages, indicating that color words from both languages created response conflict, contrary to a strict response set effect. A sustained negativity (SN) followed with larger amplitude for color incongruent than congruent trials, resolving earlier for between than within language Stroop. This effect shared timing (550–700 ms), but not morphology or scalp distribution with the commonly reported sustained potential. Finally, larger negative amplitude (200–350 ms) was observed between than within languages independent of color congruence. This negativity, likened to a no-go N2, may reflect processes of inhibitory control that facilitate the resolution of conflict at the SN, while the N450 reflects parallel processing of distracter words, independent of response set (or language). In sum, the BWLS reflects brain activity over time with contributions from language and color conflict at different points.

Keywords: bilingual, Stroop, response conflict, between language interference, N450, N2, event related potential, language dominance

INTRODUCTION

The Stroop effect has captivated researchers for over 75 years and has resulted in a vast (and daunting) body of literature. Versions of the Stroop paradigm have been used to study diverse cognitive phenomena, like selective attention, inhibition and executive control, conflict detection and monitoring, and automaticity and lexical access (see MacLeod, 1991), and have been used clinically to test for deficits in many areas (Green et al., 2010; Peckham et al., 2010; Pukrop and Klosterkötter, 2010). In the field of bilingualism, the Stroop paradigm has been commonly used to analyze the degree of interference or alternatively the degree of automaticity of access to words in each language and across languages (see Francis, 1999, for a review). The color word Stroop task (Stroop, 1935) has participants name the color of words printed in congruent (RED in red) or incongruent ink color (RED in green). The Stroop effect occurs when incongruent items elicit slower naming times than congruent items, which is generally thought to reflect interference due to the automaticity of reading words compared to naming colors. Bilinguals add the complexity of being able to perform the Stroop task in both of their languages. Moreover, the languages used for the distracter words and naming can match (within) or not (between), such that interference within each language and between languages can be measured. Because the Stroop paradigm taps into a complex set of cognitive processes, there is still much debate over the nature of this powerful effect. The goal of the current study is to examine the behavioral and neural correlates

of the bilingual Stroop task to inform word access, attention, and inhibition in the bilingual brain, as well as the nature of the Stroop effect more generally.

The Stroop effect has commonly been explained as a response level conflict, by accounts like the relative speed of processing – where competition occurs strictly at response, in having to choose the color over the faster processed word – and automaticity of access – where faster spread of activation throughout a network of concepts, and inversely smaller attentional demands, occurs for more automatic processes, like reading than naming (see MacLeod, 1991). Connectionist models of the Stroop, such as Cohen et al.'s (1990) model propose that interference can arise from any level of processing, from input to output. Information from the color and the word are processed in parallel in a distributed network with interconnections that are weighted based on experience. Attention plays a critical role in tuning these weights, such that an attentional set can be created for the specific task and even the specific response set simply by virtue of the strength of the connections between the attended items. MacLeod (1991; MacLeod and MacDonald, 2000) has argued that connectionist models present a more parsimonious account of the many factors that affect performance on Stroop tasks, accounting for both the speed of processing and automaticity differences. However, these models do not fully address the nature of the bilingual Stroop.

The Stroop effect is modulated by factors unique to operating in a bilingual mode. There is even some evidence that bilinguals can perform better on the Stroop task compared to monolinguals (Bialystok et al., 2008), a skill thought to emerge from the cognitive demands of managing two languages. Individual factors, such as dominance and relative proficiency in the languages (Mägiste, 1985; Chen and Ho, 1986; Tzelgov et al., 1990; Francis, 1999; Rosselli et al., 2002; Zied et al., 2004; Gasquoine et al., 2007), and form level factors of the stimuli, such as orthographic or phonological overlap between the languages (Preston and Lambert, 1969; Roelofs, 2003), both affect performance on the Stroop task. Bilinguals with one dominant language (herein, unbalanced bilinguals) experience greater Stroop interference when performing in the dominant than weaker language on within language trials, and experience more interference from distracter words written in the dominant than the weaker language on between language trials. In contrast, bilinguals with equivalent proficiency in both languages (herein, balanced bilinguals) generally exhibit no difference in the amount of interference across their languages, both naming within or between languages. This dynamic has been shown to change as the relative proficiency of a bilingual's languages changes (Mägiste, 1984, 1985; Chen and Ho, 1986).

In addition, bilinguals experience different magnitude of Stroop interference based on the degree of overlap of the word forms across languages (Sumiya and Healy, 2004). When color words share orthographic features across languages (green, *grun*) the magnitude of the Stroop effect is equivalent within a language (written and naming languages are the same) and between languages (Roelofs, 2003). However, when there is no orthographic overlap across languages (black, *schwarz*) the within language Stroop effect (incongruent versus congruent) is on average twice the magnitude of the between language effect (Francis, 1999). This has been referred to recently as the within language Stroop superiority effect (WLSSE; Goldfarb and Tzelgov, 2007), but we feel this inappropriately deemphasizes the importance of the between language effect. Therefore, we refer to this between-within language Stroop difference herein as the BWLS or the bilingual Stroop effect, interchangeably. This phenomenon was first observed by Dalrymple-Alford (1968), Dyer (1971) and Preston and Lambert (1969) and has since been replicated across several languages and tasks (Dyer, 1971; Chen and Ho, 1986; Tzelgov et al., 1990; Goldfarb and Tzelgov, 2007; see reviews by MacLeod, 1991; Francis, 1999). Spanish and English bilinguals (our target sample) generally show this BWLS (Preston and Lambert, 1969; Dyer, 1971), with few exceptions (Rosselli et al., 2002).

Under the accounts of the Stroop effect discussed above, which do not directly address the bilingual language system, it is clear how the proficiency of a language could affect the automaticity and/or speed of processing of the words in each language, but it is not clear how within language distracters elicit a significantly larger effect than between language distracters without further restrictions on the processors. This complexity is a result of bilinguals having two lexical representations for a single concept ("red" and "rojo" for concept RED Okuniewska, 2007). There is growing support for a model of bilingual lexical access in which both languages are non-selectively activated, at least at some stages of word recognition, even if processing demand is restricted to one language (Green, 1998; Spivey and Marian, 1999; Dijkstra and Van Heuven, 2002; Rodriguez-Fornells et al., 2005; Costa et al., 2006; Sunderman and Kroll, 2006). These lexical items must be kept at bay when they are not needed, but there is less of a consensus about how bilinguals, particularly those with high proficiency in a second language, prevent cross language interference.

Some contend that a mechanism of inhibition is required (Green, 1998; Kroll et al., 2010), while others propose that only language relevant items are "flagged" when attending to one language on a task, creating an attentional set of plausible responses (Roelofs, 2003, 2010). A third account proposes a mechanism of access through activation thresholds similar to other connectionist models (Dijkstra and Van Heuven, 2002). Spread of activation can occur between languages at various levels of processing, from semantic (Dijkstra et al., 1998; Lemhöfer and Dijkstra, 2004) to orthographic (Dijkstra et al., 1998; Jared and Kroll, 2001), and as a function of proficiency (see also Sunderman and Kroll, 2006, for a different account). Only one of these models has addressed the BWLS directly, claiming that it is something equivalent to a response set effect in monolinguals (Roelofs, 2003, 2010; Goldfarb and Tzelgov, 2007).

A response set effect (or membership effect) is observed when distracter words that are actively used for responding on the task, e.g., GREEN, RED, YELLOW, BLUE, cause more interference (larger Stroop effect) than other color words that are not being actively used to respond, e.g., PINK (Klein, 1964; Proctor, 1978; Glaser and Glaser, 1989; Lamers et al., 2010). Most accounts of the response set effect propose that it occurs at response and not during access to meaning. Cohen et al. (1990) describe response set effects as occurring at the output level of processing by attentional selection of a set of relevant responses. In a slightly different account, Roelofs (2003, 2010) restricts the response set effect to the response level, but does so by "flagging" the response relevant items

at the conceptual level in the multi-tiered WEAVER++ model. The flag results in setting and maintaining an attentional set for the response relevant items (see also Treisman and Fearnley, 1969), shielding valid responses from interference anywhere except at the output layer (response selection). Hence, response set effects elicit response conflict, not because the response-irrelevant words elicit competing responses directly, but rather by spread of activation to the response set at the semantic level. It has been argued that this attentional set account can better explain the response set effect than models that propose inhibition of irrelevant responses during stimulus evaluation (see Lamers et al., 2010). Roelofs has argued that the BWLS can be explained parsimoniously with monolingual data as a response set effect. Similar to the word PINK in the example above, the between language words, that is words that are viewed but not actively prepared for naming, e.g., VERDE, ROJO, AMARILLO, AZUL, receive less activation than the equivalent within language response set of words. In this way, the BWLS effect would be caused by differential spread of activation from the response set to related color words in the other language. If this is the case, then there should always be greater activation for response set items, and color incongruent items should be named more slowly for the response relevant than irrelevant language. Similarly, the neural correlate for the BWLS should reflect this differential spread of activation, perhaps as a modulation of amplitude from response relevant to irrelevant but related words.

This is the first study to use event related potentials (ERP) to address the source of the BWLS. In recent history, the debate over the source of Stroop interference, more generally, has been informed by electrophysiological techniques, which provide a way of experimentally disentangling semantic and response level effects. Scalp-recorded ERP, which have extraordinary temporal resolution (on the order of milliseconds), are especially well suited to investigate the timing of cognitive events. Early ERP studies of Stroop interference focused on the P300 –a component found to vary in latency with stimulus evaluation, but not response selection (Kutas et al., 1977; for a review of the P300, see Polich, 2007). Since the P300 latency is insensitive to color congruence on the Stroop task, the Stroop effect must occur later in processing, that is at response selection (Duncan-Johnson and Kopell, 1981; Ilan and Polich, 1999; Rosenfeld and Skogsberg, 2006; however Lansbergen and Kenemans, 2008, found modulation of P300 with low probability of Stroop trials).

In fact robust Stroop effects have been observed later in time at the N450 (or medial frontal negativity – MFN) and the conflict sustained potential or SP (Rebai et al., 1997; West and Alain, 1999; Liotti et al., 2000; Markela-Lerenc et al., 2004; West et al., 2004, 2005; Larson et al., 2009). While the functional significance of these components is not yet fully understood, they are thought to index different levels of conflict processing and are distinguished both by what modulates them and topographical distribution. The conflict SP, which can range in latency and duration based on task demands, generally occurs after the N450, showing increased amplitude for color incongruent than congruent trials (West and Alain, 1999; Liotti et al., 2000; West, 2003; Markela-Lerenc et al., 2004; West et al., 2005; Larson et al., 2009). The activity in this window may reflect a complex of cognitive processes, including response selection, and response monitoring and conflict adaptation, respectively by region of the SP (West et al., 2005; Chen et al., 2011).

The N450 precedes the SP as a medial fronto-central negativity between 300 and 500 ms post-stimulus onset. It is more negative in amplitude for color incongruent than color congruent stimuli, and increasing the degree of conflict increases N450 amplitude (West and Alain, 2000). Though its timing can vary with task demand, the N450 has been observed on a variety of Stroop-like tasks (West et al., 2005), with both covert (silent naming) and overt (naming aloud) responses (Liotti et al., 2000). The component's neural generators have been source localized to the anterior cingulated cortex (ACC; West, 2003; Markela-Lerenc et al., 2004). Some have argued that the ACC is responsible for "directing attention to a goal, even in the absence of conflict" (MacLeod and MacDonald, 2000), while others contend that it is responsible for conflict detection and monitoring (Van Veen and Carter, 2002; Carter and Van Veen, 2007) and that separate parts of the ACC respond to semantic (stimulus) and response conflict (Roelofs, 2003; van Veen and Carter, 2005; Wendt et al., 2007; Aarts et al., 2009; Bialystok and Craik, 2010). At least one study suggests that the ACC should be more involved in between- than within language processes (Abutalebi et al., 2008) to prevent interference from the non-target language.

The N450 effect has been observed for both response and non-response type conflict on a counting task, suggesting that it might be sensitive to both incongruent but response eligible (i.e., response set) and incongruent but response ineligible items (West et al., 2004). This would suggest that both within and between language words might modulate N450 amplitude. However, a more recent study showed that only response conflict, and not stimulus conflict, modulated the N450 on a 2-1 mapping color word Stroop task (Chen et al., 2011). By mapping two color words to one finger (index finger, BLUE/GRAY; middle finger, GREEN/WHITE; ring finger, YELLOW/PURPLE), the source of conflict was parsed by presenting trials with color incongruent words that created stimulus (GREEN/WHITE) or response (and stimulus) conflict (YELLOW/GRAY; Chen et al., 2011). N450 amplitude was more negative for response incongruent than color congruent trials, but no different for stimulus incongruent and congruent trials. Based on these findings, the BWLS may be reflected as a modulation of the N450, with a larger Stroop effect for between than within language trials.

Finally, response set (and the BWLS) may modulate earlier ERP components than the N450 and conflict SP, in particular the N2 (Folstein and Van Petten, 2008). Although the conflict N2 has not been robustly elicited in a Stroop task (West et al., 2005), its amplitude increases with increasing magnitude of conflict on other tasks, like the Eriksen flanker task (Van Veen and Carter, 2002; Wendt et al., 2007). If the conflict N2 is sensitive to the degree of conflict on the bilingual Stroop task, then greater N2 amplitude might be expected for within than between language distracters. Alternatively, attention to response relevant information, or attentional set, specifically in word recognition tasks, has been shown to modulate N2 (or N200) amplitude with increased negativity for attention to orthographic features of a word (Ruz and Nobre, 2008; see also Grainger et al., 2006, for a similar component that is modulated by orthographic processes in a priming paradigm). The N2

has been modulated on bilingual tasks that focus attention on one language at a time or cause a switch between languages (Jackson et al., 2001; Rodriguez-Fornells et al., 2005). In addition, Proverbio et al. (2009) found that bilinguals can use orthographic information to distinguish between real and pseudo native language words (Italian) as early as 160–180 ms. Hence, the language of response relevant words in the bilingual Stroop task may be detected and processed early, reflected by modulation of the N2 (see Atkinson et al., 2003, for early perceptual effects in a Stroop task).

The current study used behavioral and electrophysiological measures to investigate how Spanish–English bilinguals process language and color congruence in a modified bilingual Stroop task across two experiments. Our central aims were to investigate (1) the unique contribution of language incongruence in the bilingual Stroop paradigm and (2) the temporal dynamics and neural correlates of cognitive control in balanced bilinguals while performing a bilingual Stroop task. In Experiment 1, we collected response time (RT) and error data across single and mixed language blocks to determine the pattern of within and between language effects for our sample (Spanish–English bilinguals) and to explore the possibility that balanced and unbalanced bilinguals use different strategies in mixed versus single language context to manage cross language interference. In Experiment 2, we collected ERP data using EEG to record brain activity while balanced bilinguals performed the single language blocks from Experiment 1 both overtly (for behavioral analysis) and covertly (for ERP analysis) to determine the source of the bilingual Stroop effect or BWLS.

PART I

EXPERIMENT 1

The primary goal of Experiment 1 was to determine the pattern of within- and between language Stroop effects in our sample population of Spanish–English bilinguals. We manipulated several variables that had been tested separately in previous studies to attempt to create a complete picture within the same individuals. First, researchers have been inconsistent in their method of categorizing their study population, which may account for the variability in observing the BWLS across studies (e.g., Rosselli et al., 2002). Here we use a battery of independent measures to categorize our participants into separate groups, as proficient balanced bilinguals and bilinguals with a dominant language. Based on previous findings, we expected to observe a BWLS for both groups, but predicted that language dominance would play a role in the size of the BWLS, with larger effects when reading the dominant than non-dominant language (Dyer, 1971). Alternatively, balanced bilinguals might not show a BWLS effect if the strength of the connections for words is equivalent between and within languages. Second, previous research has shown that performance can be affected by the presence of two language simultaneously (mixed language blocks) compared to processing a single language (e.g., Christoffels et al., 2007). This may be due to the specific strategy adopted to cope with each stimulus type. We included both mixed and single language blocks to test the robustness of the BWLS. We predicted that the BWLS would be observed for both types of stimuli, but that the nature of the BWLS could vary. Specifically, interference in the form of slower RTs would be smaller during single than mixed language blocks, since the distracter language could be consistently inhibited. Finally, if the BWLS is the equivalent of a response set effect in monolinguals then color-naming times should always be slower for within language than between language trials.

Methods

Participants. Ninety-two Spanish–English bilinguals, recruited from the University of Texas at San Antonio (UTSA) and the University of Texas Health Science Center San Antonio (UTHSCSA) were paid for their participation. Data was excluded for 6 participants due to experimenter error or equipment failure and 12 participants as outliers (±2 SD from the mean) based on RT (4), accuracy (2), language dominance (4), or age[1] (2). The remaining 74 participants (mean age 25.88 years, SD = 6.56, range = 18–46 years, and handedness: right = 70, left = 4) included 50 women and 24 men, 68 (91.9%) of which reported being of Hispanic origin. All participants had normal or corrected-to-normal vision and reported no cognitive or physical impairments that could affect their performance on the task.

Language profiles. A total of 12 verbal fluency tests (VFT) were used to screen potential participants by phone; 1 min was given per test to name as many words as possible beginning with F, A, or S for English and P, T, or M for Spanish, or that fit into the categories of fruits, vegetables, or animals in each language. Proper names, repetition and variations of the same word were excluded; the number of remaining words were averaged for each language separately. Individuals with a minimum five-word average in the non-dominant language were subsequently tested on-site with a series of language measures. The 60-picture Boston naming test (BNT: Kaplan et al., 2001) was administered untimed in one language then the other. The order of languages tested on the VFT and BNT was counterbalanced across participants. The language history questionnaire (LHQ) assessed, for each language, the age of exposure, percent daily use and self-assessed ability in reading, writing, comprehension, and listening (measured on a scale of 1–7 with "beginner" at 1, "intermediate" at 4, and "native speaker" at 7). Finally, word-reading (color words in black font) and color-naming (color circles) times were measured in each language (random order per participant; 1 40-trial block for each task/language with 10 presentations of each item). In addition to the language battery, participants completed a biographical questionnaire (e.g., age, ethnicity, and hearing and sight conditions) and an abridged version of the Edinburgh Handedness Inventory.

Boston naming test scores and reading and naming times were used as objective productive-language measures to group participants as balanced ($N = 24$) or unbalanced bilinguals ($N = 50$)[2]. Participants were operationally defined as balance bilinguals if they had at least two of the three following language scores: (1) a non-significant difference (t-test, $p < 0.05$) between Spanish and English reading times or (2) naming times and (3) a difference of 10 points or less between their Spanish and English BNT scores. Unbalanced bilinguals performed better (i.e., faster, more

[1] Participants excluded for age were done so based on findings that indicate Stroop performance declines after age 55 (Jolles et al., 1995).

[2] Performance on the VFT and BNT were highly correlated [$r(87) = 0.80, p < 0.01$].

accurately and named more pictures) in the same language on at least two of the three measures[3]. **Table 1** shows performance on the language measures for each group.

[3]One participant was included as balanced having scored as English dominant on one measure, Spanish dominant on another and balanced on the third, resulting in no clearly dominant language. This participant tested as balanced on two of the three measures upon retesting the naming and reading time measures for participation in Experiment 2. This occurred with other participants as well, who switched from dominant in one language to balanced in both, or vice versa, on a specific measure. This highlights the dynamic nature of bilinguals over time, and the importance of collecting more than one measure of language proficiency/dominance, in particular when classifying individuals as balanced.

Materials and procedure. Qualified participants read and signed a consent form under the guidelines of UTSA's and UTHSCSA's Institutional Review Boards for Human Subject Research, after which they sat approximately 55″ away from a 19″ color CRT monitor and named the font color of capitalized centered half-inch tall color words (GREEN, BLUE, YELLOW, RED, VERDE, AZUL, AMARILLO, ROJO). Each color word appeared equally in each of the four font colors (green, blue, yellow, red). Stimuli were randomized and presented on a light gray background using E-Prime software (Psychological Software Tools, Inc., Pittsburgh, PA, USA). Each trial started with the presentation of three fixation crosses ("+++"; randomly 500–750 ms duration, with

Table 1 | Language profile means (SD), for balanced (Experiments 1 and 2, $N = 24$) and unbalanced (Experiment 1, $N = 50$) bilinguals.

Bilingual group	**Balanced (BB)**		**Unbalanced (UB)**
Experiment	**1**	**2**	**1**
Word-reading times (ms)			
English (BB)/dominant (UB)	156.49 (92.20)	135.74 (97.70)	95.98 (50.92)
Spanish (BB)/non-dominant (UB)	140.06 (81.52)	156.77 (92.89)	143.47 (67.84)
Difference	16.43 (44.07)	21.03 (63.51)	47.48 (54.14)**
Color-naming times (ms)			
English (BB)/dominant (UB)	236.70 (83.97)	247.98 (88.38)	177.99 (60.50)
Spanish (BB)/non-dominant (UB)	221.66 (85.26)	206.14 (87.27)	251.71 (83.86)
Difference	15.04 (28.02)*	41.84 (57.30)*	73.71 (59.30)**
Boston naming test (BNT)			
English (BB)/dominant (UB)	44.67 (5.84)	45.33 (5.89)	48.64 (6.92)
Spanish (BB)/non-dominant (UB)	43.08 (8.10)	43.13 (7.58)	32.52 (11.45)
Difference	2.21 (9.43)	−1.04 (8.61)	16.12 (15.09)**
Verbal fluency test			
English (BB)/dominant (UB)	13.36 (2.74)	14.56 (3.55)	14.92 (2.76)
Spanish (BB)/non-dominant (UB)	13.98 (3.27)	14.79 (3.58	11.98 (2.91)
Difference	−0.19 (2.53)	−1.10 (2.56)	2.93 (3.30)**
Percentage of daily use			
English (BB)/dominant (UB)	54.38% (20.92)	63.04% (19.53)	61.76% (23.25)
Spanish (BB)/non-dominant (UB)	44.79% (21.34)	36.96% (19.53)	38.18% (23.21)
Age of exposure			
English	6.25 years (4.91)	6.71 years (5.30)	5.35 years (4.78)
Spanish	0.08 years (0.41)	0.57 years (2.71)	1.52 years (4.61)
Perceived language ability (scale of 1–7)			
English (BB)/dominant (UB)			
Speaking	6.29 (0.81)	6.21 (1.02)	6.74 (0.57)
Comprehension	6.50 (0.78)	6.17 (1.13)	6.76 (0.43)
Reading	6.42 (0.83)	6.23 (1.18)	6.76 (0.63)
Writing	6.25 (0.99)	6.08 (0.93)	6.60 (0.76)
Spanish (BB)/non-dominant (UB)			
Speaking	6.50 (0.78)	6.42 (0.83)	5.32 (1.25)
Comprehension	6.63 (0.71)	6.50 (0.83)	5.86 (1.16)
Reading	6.12 (1.36)†	6.08 (1.50)†	5.76 (1.29)
Writing	5.83 (1.52)†	5.88 (1.48)†	5.26 (1.40)†

The Boston naming test and reading and naming times were used to categorize each subject by language balance, see Section "Methods" for criteria.

**Significant difference, $p \leq 0.05$, **significant difference, $p \leq 0.001$.*

†Range of response was from 1 to 7 in these domains; among balanced bilinguals this likely reflects less formal education in Spanish.

Differences were always English minus Spanish or dominant minus non-dominant.

200 ms blank screen ISI), followed by the stimulus (150 ms duration with 200 ms blank screen ISI; per Liotti et al., 2000), then a single fixation cross ("+") which remained on the screen until a verbal response was detected by the integrated voice-key of a PST serial response box by way of an external microphone (Psychological Software Tools, Inc., Pittsburgh, PA, USA). An additional microphone and digital recorder collected verbal responses for accuracy analyses.

A total of 8 blocks were presented, consisting of 96 trials each (768 total trials). In each block, half of the words were color congruent (CC, e.g., "RED" written in red) and half were color incongruent (CI, e.g., "BLUE" written in red), see **Table 2** for sample stimuli. Naming language was held constant across an entire block with four blocks named in English, four in Spanish (naming language order was randomized per participant). Four blocks were presented in a single language (SL, two blocks of Spanish color words and two of English color words) and four in mixed languages (ML, Spanish and English color words in the same block). To manipulate language, half of the trials in mixed language blocks, and half of the blocks in single language blocks, were printed in the same language as the naming language (language congruent trials, LC), and half were not (language incongruent trials, LI). An equal number of trials were presented in each minimal contrast (e.g., ML–LC–CC versus SL–LC–CC). Each block was preceded by a short practice session that informed the participant in which language to name the font colors. The inter-block interval lasted no longer than 5 min and the entire session lasted approximately 1.5 h.

Results

Error trials and accurate RTs were analyzed for each group separately. RTs in milliseconds were measured from the onset of the visual word to detection of the voice response (Balanced Bilinguals, $M = 375.60$, SD = 94.25; Unbalanced Bilinguals, $M = 351.96$, SD = 101.25). RTs more than ±2 SD away from the condition means and all response errors (defined as wrong font color response, wrong language response, or unintelligible response) were excluded from RT analyses. For balanced bilinguals, a 2 Block Type (single language, mixed language) × 2 Naming Language (English, Spanish) × 2 Color Congruence (congruent, incongruent) × 2 Language Congruence (congruent, incongruent) repeated-measures ANOVA was used. Since unbalanced bilinguals had a known dominant language in which they were expected to perform better, and that language was not always the same across participants, we collapsed across Naming Language to create a level of Language Dominance (dominant, non-dominant) in the ANOVA design. All planned contrasts were Bonferroni adjusted for multiple comparisons. When a Color Congruence × Language Congruence interaction was found, additional paired samples *t*-tests were conducted to evaluate the Stroop effect size (color incongruent minus color congruent trials) of within and between language interference (when naming and written languages were congruent and incongruent, respectively).

Table 2 | Sample stimuli.

	Stimulus	Language congruent response (within language trails)	Language incongruent response (between language trials)
English color congruent	RED	red	rojo
English color incongruent	BLUE	red	rojo
Spanish color congruent	ROJO	rojo	red
Spanish color incongruent	AZUL	rojo	red

During the single language blocks, words appeared consistently in one language while the naming language was either congruent (within) or incongruent (between) through out. This created separate between and within language blocks. During mixed language blocks the words appeared randomly and alternately in Spanish or English, while the naming language remained constant, creating within and between language trials within each block.

Unbalanced bilinguals.

Error analyses. Overall, unbalanced bilinguals made more errors on color incongruent than congruent trials [$M = 3.5\%$, SD = 2.4% versus $M = 0.7\%$, SD = 0.7%; $F(1, 48) = 86.174$, $p < 0.001$], and more errors on language congruent than incongruent trials [LC; $M = 5.7\%$, SD = 0.8% versus LI; $M = 1.7\%$, SD = 1.3%; $F(1, 26) = 18.580$, $p < 0.001$], **Figure 1**. Although there was a significant Color Congruence effect for both Within and Between language conditions ($p < 0.001$), the effect was

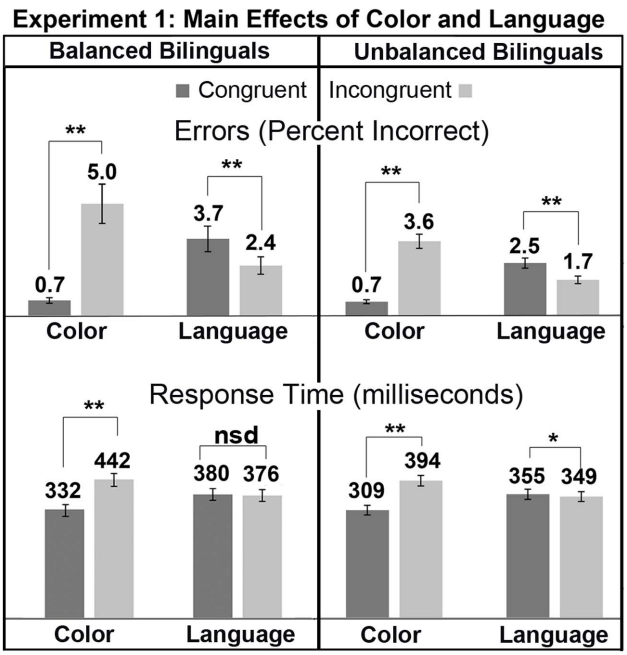


FIGURE 1 | Main effects of color congruence and language congruence from Experiment 1. Mean proportion of incorrect responses and mean response times in milliseconds reported for each group: $^{**}p \leq 0.001$; $^{*}p \leq 0.050$; nsd, non-significant differences.

significantly larger for language congruent than language incongruent trials; $F(1, 48) = 22.087$, $p = 0.0001$. Effects of Block Type and Language Dominance were not significant.

Response times analyses. Response times in milliseconds were analyzed for accurate trials only ($M = 96.43\%$, SD $= 2.33\%$). As expected, a robust Color Congruence effect was observed, with faster naming times on color congruent than incongruent trials [$M = 309.73$, SD $= 97.42$ versus $M = 394.20$, SD $= 107.27$; $F(1, 49) = 361.458$, $p < 0.001$], **Figure 1**. In addition, faster naming times were observed overall for language incongruent compared to congruent trials [$M = 348.62$, SD $= 99.54$ versus $M = 355.30$, SD $= 103.97$; $F(1, 49) = 5.185$, $p = 0.027$], and for single than mixed language trials [$M = 348.58$, SD $= 101.70$ versus $M = 355.34$, SD $= 102.25$; $F(1, 49) = 3.882$, $p = 0.054$]. These main effects were qualified by interactions between Color Congruence and Language Congruence, $F(1, 49) = 32.078$, $p < 0.001$, and Block Type: Color Congruence by Language Congruence by Block Type, $F(1, 49) = 7.173$, $p = 0.010$, and Language Congruence by Block Type, $F(1, 49) = 33.042$, $p < 0.001$, but not Color Congruence by Block Type. Analyses focusing first on the Color Congruence effect then the Language Congruence effect explain the source of these interactions.

The Color Congruence effect was observed both within and between languages ($p < 0.001$), but the effect was significantly larger (i.e., a larger difference between color congruent and incongruent trials) on language congruent (within language) than language incongruent trials [between languages; $M_{\text{diff}} = 98.97$, SD $= 43.74$ versus $M_{\text{diff}} = 69.97$, SD $= 26.76$, $t(49) = 5.664$, $p = 0.001$]. This classic between- versus within language Stroop effect difference, or BWLS, was present for both mixed- and single language presentation ($p < 0.005$), but was larger for mixed language trials [$t(49) = 2.678$, $p = 0.010$], **Figure 2**. Language congruent trials were slower than language incongruent trials only during single- ($p < 0.001$), and not mixed language presentation. Planned contrasts revealed an interesting pattern in the simple effects. The effect of Language Congruence for single language trials was carried by the color incongruent trials, **Table 3**. There was no effect of language congruence when color was congruent, but when color was incongruent, language congruent trials were significantly slower than language incongruent trials ($p < 0.001$), indicating that interference from the color incongruent distracter word was greater for the response relevant language. In contrast, for mixed language trials, there was an effect of language congruence both when color was congruent and incongruent, but the effects were opposite of each other, **Figure 2**.

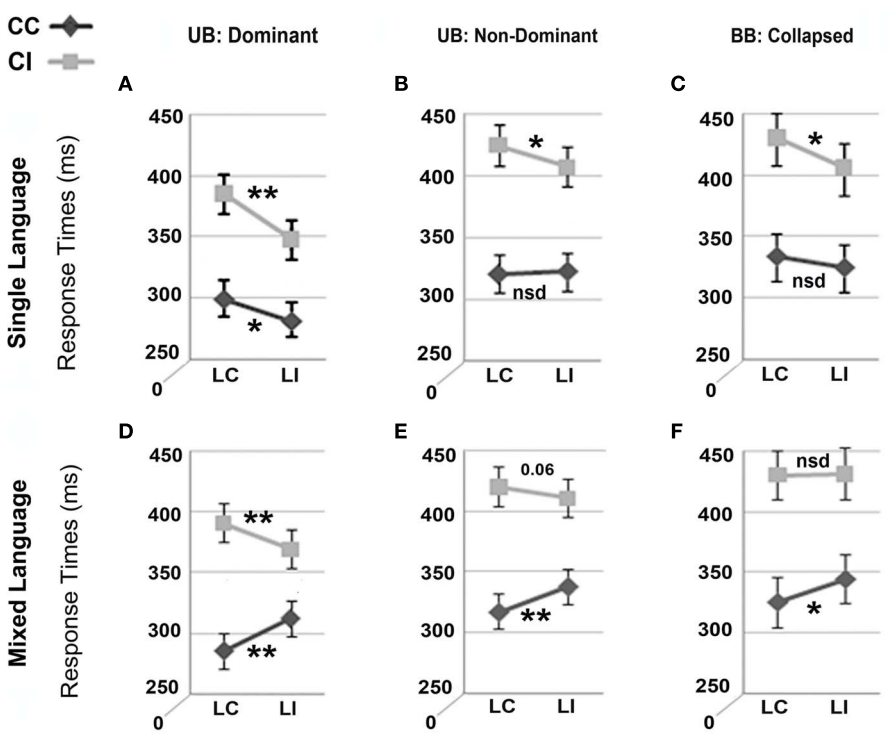


FIGURE 2 | Mean response times in milliseconds showing the interaction between color congruence and language congruence by block type and group from Experiment 1. Results are presented for unbalanced bilinguals (UB) for dominant **(A,D)** and non-dominant **(B,E)** naming languages separately and for balanced bilinguals (BB) collapsed across naming language **(C,F)**. Panels **A–C** show results for blocks of stimuli presented in a single written language, collapsed across Spanish and English; panels **D–F** show results for stimuli presented alternately in Spanish and English in the same block. In all six plots, the effect of color congruence was significant at $p \leq 0.001$ and this effect was significantly larger within than between languages at $p \leq 0.05$. All other effects noted: $^{**}p \leq 0.001$; $^{*}p \leq 0.050$; nsd, non-significant differences. CC, color congruent; CI, color incongruent; LC, language congruent; LI, language incongruent.

Table 3 | Simple effects means (SD) in milliseconds.

Bilingual group	Balanced (BB)		Unbalanced (UB)
Experiment	**1**	**2**	**1**
English (BB)/dominant language (UB) single language blocks			
Color congruent language congruent (CCLC)	349.41 (96.95)	321.90 (91.44)	299.25 (113.92)
Color incongruent language congruent (CILC)	446.88 (102.94)	409.25 (94.82)	385.05 (112.77)
Color congruent language incongruent (CCLI)	331.66 (98.32)	320.52 (102.94)	347.88 (104.79)
Color incongruent language incongruent (CILI)	415.24 (121.57)	388.76 (101.90)	377.43 (105.95)
English (BB)/dominant language (UB) mixed language blocks			
Color congruent language congruent (CCLC)	336.46 (110.95)		286.08 (101.93)
Color incongruent language congruent (CILC)	438.25 (100.87)		390.32 (114.71)
Color congruent language incongruent (CCLI)	355.24 (101.56)		312.45 (102.66)
Color incongruent language incongruent (CILI)	435.91 (107.05)		369.25 (108.78)
Spanish (BB)/non-dominant language (UB) single language blocks			
Color congruent language congruent (CCLC)	318.22 (98.96)	297.02 (105.57)	320.67 (107.78)
Color incongruent language congruent (CILC)	405.88 (111.63)	380.26 (101.11)	424.50 (117.29)
Color congruent language incongruent (CCLI)	310.80 (107.27)	285.16 (100.91)	322.86 (103.08)
Color incongruent language incongruent (CILI)	391.86 (101.45)	350.66 (100.29)	406.98 (113.27)
Spanish (BB)/non-dominant language (UB) mixed language blocks			
Color congruent language congruent (CCLC)	312.25 (97.22)		317.27 (105.28)
Color incongruent language congruent (CILC)	420.02 (102.47)		419.27 (116.78)
Color congruent language incongruent (CCLI)	326.01 (91.33)		337.79 (100.39)
Color incongruent language incongruent (CILI)	415.47 (94.84)		410.31 (112.08)

There were no mixed language blocks in the Experiment 2.

Single (Experiments 1 and 2) and mixed language (Experiment 1) blocks for balanced (Experiments 1 and 2, N = 24) and unbalanced bilinguals (Experiment 1, N = 50) in each naming language (English/Spanish).

Specifically, when color was congruent, language congruent trials were significantly *faster* than language incongruent trials (CCLC versus CCLI, $p < 0.001$), but when color was incongruent, language congruent trials were significantly *slower* than language incongruent trials (CILC versus CILI, $p < 0.001$). The language incongruent trials were slower overall during mixed than single language presentation (CCLI, $p < 0.001$; CILI, $p < 0.004$), indicating that the language of the distracter words caused more interference during mixed language presentation. The possible effect of strategy and processing of non-response set words is discussed below.

Finally, with regard to naming language, unbalanced bilinguals were faster overall when responding in their dominant than in their non-dominant language [$M = 333.97$, SD = 101.70 versus $M = 369.96$, SD = 104.46; $F(1, 49) = 43.008$, $p = 0.001$]. The effect of color congruence was modulated by language dominance [Color Congruence by Dominant Language, $F(1, 49) = 7.535$, $p = 0.008$; Color Congruence by Dominant Language by Block Type, $F(1, 49) = 4.516$, $p = 0.039$]. During mixed language presentation, the Color Congruence effect was the same whether naming in the dominant or non-dominant language; conversely, the effect of language dominance was the same for both color congruent and incongruent trials. However, during single language presentation, the Color Congruence effect was larger when naming in the non-dominant and reading the dominant language than vice versa; conversely, the difference between the dominant and non-dominant response languages was greater for color incongruent than color congruent trials [$t(49) = 3.52$, $p = 0.001$].

No other effects were significant.

Balanced bilinguals.

Error analyses. Data from 26 balanced bilinguals was included in the error analyses. One participant did not have complete accuracy data due to a voice-recording error on the last block of trails. Based on this individual's percent errors on the other blocks (4.9%), we estimate that approximately 5 error trials were not accounted for here and were included in the RT analyses.

Overall, balanced bilinguals made more errors on color incongruent than congruent trials [$M = 5.2\%$, SD = 4.45% versus $M = 0.7\%$, SD = 0.71%; $F(1, 23) = 24.311$, $p < 0.001$], and more errors on language congruent than incongruent trials [$M = 3.7\%$, SD = 3.09% versus $M = 2.4\%$, SD = 2.07%; $F(1, 23) = 13.725$, $p = 0.001$], **Figure 1**. Although there was a significant Color Congruence effect both Within and Between language conditions ($p < 0.001$), the effect was significantly larger for language congruent than language incongruent trials [$M_{diff} = 5.7\%$, SD = 1.17% versus $M_{diff} = 3.6\%$, SD = 0.74%; $F(1, 23) = 16.695$, $p < 0.001$].

There were no main effects of Naming Language or Block Type. These factors did, however, interact: Naming Language × Block Type, $F(1, 23) = 6.425$, $p = 0.019$; Block Type × Naming Language × Language Congruence, $F(1, 23) = 4.652$, $p = 0.042$. These effects are consistent with a speed–accuracy trade off when naming in Spanish (see RTs below).

Response time analyses. Response times in milliseconds were analyzed for accurate trials only ($M = 95.39\%$ of total trials, SD = 3.21%; see text footnote 4). As with unbalanced bilinguals, balanced bilinguals showed a robust effect of Color Congruence, with faster naming times on color congruent than incongruent trials [$M = 332.08$, SD = 94.22 versus $M = 424.13$, SD = 98.95; $F(1, 23) = 289.33$, $p = 0.001$]. There was no main effect of Language Congruence, but Color Congruence and Language Congruence interacted, $F(1, 23) = 14.257$, $p = 0.001$. As with the error data, although a Color Congruence effect was observed both within and between languages ($p < 0.001$), the effect was larger on language congruent (within language) than incongruent trials [between languages; $M_{diff} = 100.50$, SD = 29.04 versus $M_{diff} = 83.60$, SD = 28.33; $t(23) = 3.776$, $p = 0.001$], see **Table 3** and **Figure 2**.

There was no main effect of Block Type, and no interaction between Block Type and Color Congruence, or Block Type, Color Congruence, and Language Congruence, indicating that, contrary to unbalanced bilinguals, this within- versus between language difference on the color congruence effect was not larger during mixed- than single language presentation, **Figure 2**.

However, similar to unbalanced bilinguals, a Block Type by Language Congruence interaction revealed a trend for faster naming times on language incongruent than congruent items [$M = 365.51$, SD = 103.34 versus $M = 382.21$, SD = 93.29; $F(1, 23) = 9.693$, $p = 0.005$] on single language trials; language incongruent took longer than language congruent items on mixed language trials ($M = 387.35$, SD = 96.53 versus $M = 377.37$, SD = 100.71; $p = 0.047$), see **Figure 2**. No other interactions with Block Type reached significance.

Although the participants were considered balanced in their two languages based on performance on the language measures (see **Table 1**), naming times were faster overall in Spanish[4] than English [$M = 358.14$, SD = 98.66 versus $M = 389.58$, SD = 100.31; $F(1, 23) = 12.423$, $p = 0.002$]. There were no significant interactions with Naming Language.

Discussion

The primary goal of Experiment 1 was to determine the pattern of within- and between language Stroop effects in our sample population of Spanish–English balanced and unbalanced bilinguals. In brief, we observed the classic Stroop effect, with longer RTs for color incongruent than congruent trials. This effect was observed both when the naming and reading languages were the same (within language) and when they were different (between language). In addition, we observed a larger Stroop effect within than between languages –the bilingual Stroop effect or BWLS, which was present across all conditions, regardless of group, block type or naming language (**Figure 2**). We discuss the BWLS effect in detail, beginning with naming language and block type effects for each group separately.

The pattern of Stroop effects was very similar for both groups of bilinguals. The primary difference between the groups was a larger Stroop effect for unbalanced bilinguals when naming in the non-dominant language – showing more cross language interference from reading the dominant than non-dominant language. Balanced bilinguals showed the same pattern in both languages. These findings are consistent with previous research (Dyer, 1971) and can be explained by a difference in automaticity of access to the words in each language based on dominance (Cohen et al., 1990; Kroll and Stewart, 1994). Interestingly, the language dominance effect was observed only for single language blocks, and disappeared on mixed language trials. This pattern reflects a differential mixing cost across the groups driven by the distracter language. Although naming was performed in a single language in the current study, unbalanced bilinguals exhibited a mixing cost in line with Christoffels et al. (2007), who observed mixing costs for German–Dutch unbalanced bilinguals on a picture-naming task, with longer RTs for mixed than single language trials. Perhaps the language dominance effect disappears in unbalanced bilinguals, because they experience more interference when naming between languages on mixed language trials, where reading both languages prevents one from becoming fully active as in the single language case.

Bilingual word recognition models, such as BIA+ (Dijkstra et al., 1998; Green, 1998; Dijkstra and Van Heuven, 2002), assume that some form of inhibition is required to allow one language to surface as the target (for an alternative view see the WEAVER++ model, Roelofs, 2003, 2010; Lamers et al., 2010). For bilinguals with asymmetric language dominance, stronger inhibition is required to keep the dominant language in check when operating in the weaker language, which in turn requires more effort to overcome in order to access the dominant language again. During single language presentation, the need to inhibit the distracter words on between language trials presents an asymmetric problem biased toward more interference from the distracters when naming in the non-dominant language. However, during mixed language presentation, the need to inhibit distracters from the stronger language is present both when naming in the dominant and non-dominant languages. Thus, the powerful effect of language dominance disappears when the languages are presented together.

An alternative explanation for the slower naming times on mixed than single language trials could be a cost from switching languages from trial to trial, in line with the idea that a language switch reverses activation and inhibition patterns in the languages (e.g., BIA+ or Green Inhibitory control model; Jackson et al., 2001; Moreno et al., 2002; Hernandez, 2009; Midgley et al., 2009). However, analyses of variance showed no difference in naming times between switch and non-switch trials in the mixed language blocks for either group, and switching did not interact with response language (no switch-cost asymmetry). Hence, the difference in the Stroop effect between mixed and single language blocks may be due to the mere presence of both languages, rather than switching costs *per se*. Activation and inhibition of the non-target language will be tested further in Experiment 2.

Despite these group differences, the presence of a between language Stroop effect across all conditions (groups, blocks, naming

[4]Balanced bilinguals as a group (but not all individuals) were faster at naming colors in Spanish than English on the baseline color-naming task, paired samples $t(26) = 2.768$, $p = 0.010$ (**Table 1**). However, unbalanced bilinguals named colors in their dominant language equally fast whether they were dominant in English or Spanish, and is therefore not due to a general naming bias for Spanish as a language (c.f., Chen and Ho, 1986).

language) indicates that the words from the non-target language consistently cause interference, in line with our bilinguals performing in a "bilingual mode" (Grosjean, 1998) and contrary to findings that bilinguals can ignore the irrelevant language (Rodriguez-Fornells et al., 2002). The second and key finding from Experiment 1 was the presence of the bilingual Stroop effect or BWLS across all conditions. As discussed above, it has been proposed that the BWLS is simply a response set effect, equivalent to the effect observed in monolinguals (Roelofs, 2003, 2010; Goldfarb and Tzelgov, 2007). Bilinguals are thought to treat the color words in the other language as response-irrelevant, similar to irrelevant words in the same language, because they are not actively producing those words on a given block of trials. The BWLS arises from response conflict, but the source of the conflict may arise at output or at higher levels of processing (Cohen et al., 1990; Roelofs, 2003, 2010). To look for response set effects, it was necessary to look at the Stroop data in an unconventional way; rather than look for color-Stroop effects across languages, we looked at the effect of language in the presence or absence of color interference.

Figure 2 shows that although there was a BWLS in all conditions, the exact pattern of effects varied within each group differently by block type and naming language. This pattern provides only partial support for the response set explanation, where 2 things should be true. First, color congruent items should be named fastest for the response relevant than irrelevant language, due to the converging information in the color and word channels. This was observed consistently during mixed language presentation, regardless of language dominance (**Figures 2D–F**), indicating that the language of the distracter word can elicit naming interference in the absence of color interference (i.e., the word BLUE in blue versus the word AZUL in blue). However, this was not true during single language presentation (**Figures 2A–C**). In the absence of color-Stroop interference (color congruent trials – CC) there was an effect of language congruence only for unbalanced bilinguals when naming in their dominant language (**Figure 2A**). In this case, language congruent items were named *slower* than language incongruent items[5]. This interaction indicates that during mixed language presentation, the language of the distracter word can elicit interference in the absence of color interference (i.e., the word BLUE in blue versus the word *AZUL* in blue), which argues against a simple response set effect (Roelofs, 2003; Goldfarb and Tzelgov, 2007) or that the task-irrelevant language can be ignored (Rodriguez-Fornells et al., 2002). This may be due to the strength of the connections for the weaker language (e.g., Cohen et al., 1990), such that even processing a fully congruent word in the dominant language leads to slower color-naming times compared to reading a weaker cross language equivalent. However, the fact that there was no difference between language congruent and incongruent items for balanced and unbalanced bilinguals reading their dominant language (**Figures 2B,C**), indicates that response set did not play a role on color congruent trials. Overall, these effects suggest that bilinguals are able to control interference from the irrelevant language during single language presentation, perhaps through inhibitory mechanisms, but do less well when distracters are presented in both languages.

Second, if the BWLS is a response set effect then color incongruent items should be named slower for the response relevant than irrelevant language. This was true during single language presentation (**Figures 2A–C**), where there was consistently more interference from within language distracters (CILC) than between language distracters (CILI) regardless of naming language and in both groups. However, during mixed language presentation this difference was present only for unbalanced bilinguals naming in the dominant language (**Figure 2A**) and marginal (**Figure 2B**) or absent (**Figure 2C**) when reading a proficient language. In particular, for balanced bilinguals the source of the BWLS during single language presentation was greater interference within than between languages on color incongruent trials, but during mixed language presentation was caused by a language effect on color *congruent* trials and the absence of a language congruence effect on color incongruent trials. Therefore, although the magnitude of the BWLS was the same across blocks, the cause of the BWLS appears to be quite different. This may again indicate that the mere presence of both languages on mixed language blocks makes inhibiting words from the non-target language more difficult.

In brief, the results from Experiment 1 indicate that both balanced and unbalanced bilinguals were unable to ignore the task-irrelevant language (Rodriguez-Fornells et al., 2002), and that a simple response set effect does not fully account for the BWLS (e.g., Roelofs, 2003; Goldfarb and Tzelgov, 2007). The goal of Experiment 2 was to identify the electrophysiological correlates for the bilingual Stroop task in order to delineate what type of activity drives the BWLS, and the Stroop effect more generally, and at what stage of processing it occurs.

PART II: ELECTROPHYSIOLOGICAL CORRELATES FOR THE BILINGUAL STROOP EFFECT

EXPERIMENT 2

Experiment 2 was designed to uncover the cognitive and neural correlates of the bilingual Stroop effect. To make this initial ERP analysis of the BWLS feasible, we chose to begin exploring this question with balanced bilinguals during single language presentation, given that language dominance in the unbalanced bilinguals played a role in both the language and color congruence effects, and to isolate the BWLS effect in the absence of any mixing effects. Future studies are planned to explore the nature of the mixing effect and the effect of language dominance on the ERP BWLS. Thus, ERPs were recorded while balanced bilinguals performed the single language bilingual Stroop task from Experiment 1, naming the colors of color words first overtly then covertly. RT and accuracy from overt naming trials and ERPs from covert naming trails are presented herein.

The monolingual ERP literature does not provide clear predictions for the ERP correlates of the BWLS, and often do not align with the debate over the source of the BWLS in the behavioral literature. However, we predicted that, consistent with the

[5]Our color-naming baseline produced faster naming times than all other trials. Future studies could employ an improved neutral baseline to determine if this difference is facilitatory for within language or inhibitory for between language trials.

monolingual ERP Stroop literature, color congruence would modulate the N450 (Liotti et al., 2000; West et al., 2004, 2005; Chen et al., 2011). Based on the assumption that the N450 reflects response conflict, it would be present for within but not between language trials. The N2, which indexes response inhibition on both non-language (Liotti et al., 2007; Pliszka et al., 2007) and language tasks (Jackson et al., 2001; Rodriguez-Fornells et al., 2005) would likely show more negative amplitude for language incongruent than congruent trials. Finally, since the late SP is thought to index general conflict reprocessing (West, 2003) we predicted both color and language congruence effects on this component.

Methods

Participants. Participants were recruited from the UTSA and UTHSCSA general populations. Screening procedures were the same as for balanced bilinguals in Experiment 1 (see **Table 1**). Thirty Spanish–English right-handed balanced bilinguals were paid for their participation. Data from 6 participants were excluded due to excessive EEG artifact (4), recording error (1), or task performance error (1). The remaining 24 participants (age range 18–35 years; $M = 25$ years, $SD = 4.76$) included 21 women and 3 men, all reportedly of Hispanic origin. Twelve participants (50%) previously participated in Experiment 1. Inclusion criteria on the language measures were the same as for balanced bilinguals in Experiment 1. All participants had normal or corrected-to-normal vision and reported no cognitive or physical impairments that could affect task performance.

Materials and procedure. The stimuli and paradigm were similar to Experiment 1 for the single language blocks only, with a few methodological changes. First, naming on the critical ERP trials was silent (covert). Second, two measures were used to ensure naming language and performance accuracy. An overt naming block preceded each covert naming block in the same language, and eight probe trials were included in the covert blocks. These trials were underlined color words cuing the participant to name that trial aloud. Third, the fixation cross that appeared after each word remained on the screen for 1000 ms before the onset of the next trial, see **Figure 3**. Participants were asked to refrain from blinking during this time to avoid eye movement artifact in the EEG.

As in Experiment 1, the covert naming trials consisted of four single language blocks, two in Spanish and two in English (language order was randomized across subjects), for a total of 384 critical trials (equal number of randomly presented trials per condition and color in each block). An E-Prime coding error occurred that resulted in a loss of 4 trials of CCLI and 12 trials of CILI when naming in Spanish, thus, pairwise analyses of conditions were performed with trials collapsed across English and Spanish. For each language, 1 block was named in the same language as the written words (language congruent) and 1 block in the incongruent language.

Participants read and signed a consent form under the guidelines of the UTSA and UTHSCSA Institutional Review Board for Human Subject Research. Participants were fitted with EEG electrodes and sat in a sound attenuating, RF shielded chamber approximately 55″ away from a 19″ color CRT monitor. Participants were allowed to take breaks between blocks; no single break lasted longer than 5 min. The entire ERP session lasted approximately 2.5 h.

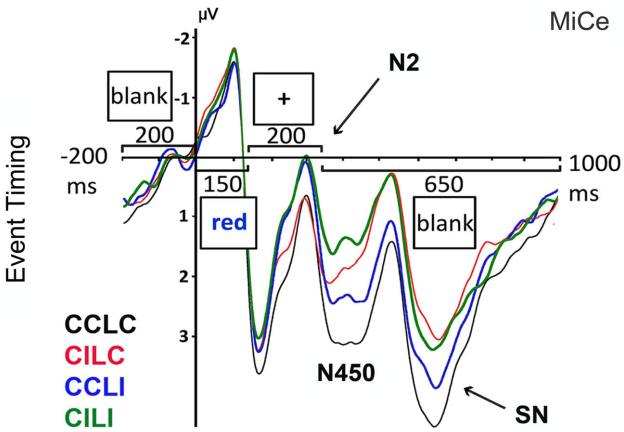


FIGURE 3 | The timing of paradigm events in Experiment 2, overlaid on grand average ERPs at electrode MiCe (vertex) time-locked to the onset of the visually presented words.

EEG recording. Continuous scalp-recorded EEG was acquired using a geodesic array of 26 pre-amplified sintered Ag–AgCl electrodes embedded in a custom electrode cap (Electro-Cap International Inc.). Additional electrodes were placed below and at the outer canthi of the left and right eyes to record blinks and eye movement respectively, and on the left and right mastoid processes to serve as offline reference. Preamplifiers in each electrode reduced induced noise between the electrode and the amplification/digitization system (BioSemi ActiveTwo, BioSemi B.V., Amsterdam), allowing high electrode impedances. Electrode offsets were kept below 40 mV. A first-order analog anti-aliasing filter with a half-power cutoff at 3.6 kHz was applied (see www.biosemi.com). The data were sampled at 512 Hz (2048 Hz with a decimation factor of 1/4) with a bandwidth of DC to 134 Hz, using a fifth order digital sinc filter. Each active electrode was measured online with respect to a common mode sense (CMS) active electrode producing a monopolar (non-differential) channel, and was referenced offline to the average of the left and right mastoids[6]. Data were processed using BrainVision Analyzer 2 (Brain Products GmbH, Munich). Non-causal Butterworth digital filters were applied with a low cutoff at 0.1 Hz (12 dB/oct) and high cutoff at 30.0 Hz (12 dB/oct). The EEG data were segmented in intervals of 1000 ms time-locked to stimulus onset, followed by DC local detrend for 100 ms blocks (Hennighausen et al., 1993) and baseline correction using −100 to 0 ms prestimulus.

Prefrontal channels were removed from analyses due to excessive artifacts restricted to those channels. The remaining 21 channels were processed using the following artifact rejection measures: maximum step of 75 μV/ms to capture voltage spikes, maximum amplitude difference of 150 μV/200 ms to capture signal drift,

[6]The average reference and average mastoid reference have shown equivalent results in previous studies (see Chen et al., 2011).

maximum amplitude of ±70 μV to capture blinks, and minimum amplitude difference of 0.5 μV/50 ms to capture flat lining and saccades. Only participants who retained 70% or more of the critical trials were included in the averages. The mean trials lost to artifact or error was 14.17%. Average waveforms were calculated for each condition time-locked to the onset of each word.

Results

Behavioral responses for overt naming trials. To determine the pattern of behavioral effects for the participants in Experiment 2, naming errors and RTs in milliseconds for the overt naming trails were analyzed using the same procedure as for balanced bilinguals in Experiment 1. As in Experiment 1, color incongruent trials elicited more errors than color congruent trials [$M = 5.7\%$, SD = 6.4% versus $M = 1.3\%$, SD = 2.8%; $F(1, 20) = 12.843$, $p = 0.002$], and the color-Stroop effect was larger for language congruent than language incongruent trials [$F(1, 20) = 5.091$, $p = 0.035$], see **Figure 4**.

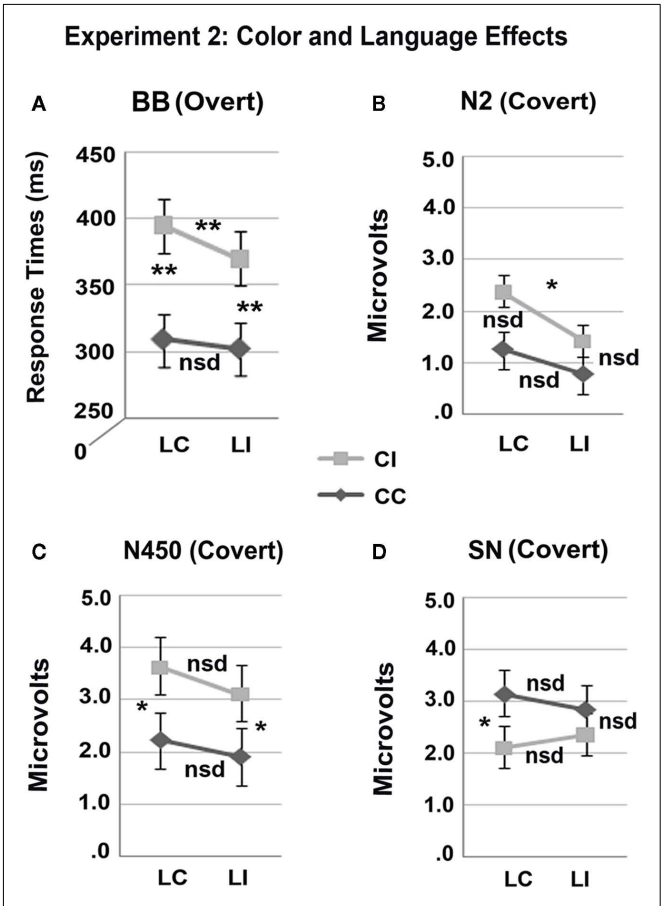


FIGURE 4 | The interaction between color and language in the overt behavioral results (mean response times) and all three times windows of the covert ERP results (mean amplitude). Note that interactions were observed only for data in **(A,D)**; only main effects of language congruence and color congruence were observed in data from **(B)** the difference between language congruent and incongruent stimuli trended at $p = 0.122$) and **(C)**, respectively: $^{**}p \leq 0.001$; $^{*}p \leq 0.050$; nsd, non-significant differences.

Similarly, slower naming times were observed for color incongruent than congruent trials, [$M = 382.23$, SD = 97.43 versus $M = 306.15$, SD = 94.60; $F(1, 23) = 149.931$, $p < 0.001$]. Unlike Experiment 1, the main effect of Language Congruence did reach significance, with faster naming times overall for language congruent than incongruent trials [$M = 352.11$, SD = 94.66 versus $M = 336.27$, SD = 97.59; $F(1, 23) = 6.004$, $p = 0.022$]. The Color Congruence effect was significantly larger within than between languages [$M_{diff} = 85.29$, SD = 37.11 versus $M_{diff} = 66.88$, SD = 30.60; $F(1, 23) = 8.840$, $p = 0.007$]. Naming times were again faster overall in Spanish than English [$M = 328.27$, SD = 99.98 versus $M = 360.11$, SD = 93.60; $F(1, 23) = 15.583$, $p = 0.001$].

Covert naming ERP results. Naming accuracy on probe trials for the covert naming blocks was at 95.4%, indicating that participants were performing the task correctly. Because responses were covert, we were unable to remove trials with naming errors. However, previous studies have shown equivalent ERP patterns from covert and overt performance on a Stroop task, supporting the validity of this task (Liotti et al., 2000). Inclusion of the few unknown error trials should not significantly affect the pattern of effects. All artifact free trials were included in the ERP analyses.

Overall, the ERP to each word was characterized by early sensory components – N1 and P2 – followed by two successive biphasic negative–positive deflections, with negative peaks at approximately 300 and 530 ms post-stimulus onset (note that the N400 that typically occurs to words is presumably suppressed due to the extensive repetition of each item), see **Figure 5**. Note that the ERP components of interest are overlaid on the visual onset and offset potentials to the fixation cross that follows the target word, see **Figure 3**. Visual inspection of the main effects of language and color congruence revealed two modulations with different timing. Language incongruent trials elicited more negativity than congruent trials starting approximately at 200 ms post-stimulus onset and ending before 500 ms, in line with the timing of the N2 (or N200) observed in the language literature, **Figure 6A**. Color incongruent trials elicited more negativity than congruent trials starting around 350 ms post-stimulus onset and resolving toward the end of the epoch, which is in line with the timing of the classic Stroop N450 in the early part of this deflection, **Figure 6B**. The effect after the N450 did not have the typical distribution or polarity shift reported in the literature for the conflict SP (e.g., West et al., 2005); hence, it is referred to herein simply as a sustained negativity (SN). However, previous findings support the disassociation of activity in these two time windows (West, 2003; Markela-Lerenc et al., 2004). Based on these contrasts three separate time windows were selected for analyses: N2 (200–350 ms), N450 (350–550 ms), and SN (550–700 ms). **Figure 4** plots the BWLS effects for mean amplitude in each time window.

Mean amplitudes for each ERP component were subjected to repeated-measures ANOVAs with Naming Language (English, Spanish) × Color Congruence (congruent, incongruent) × Language Congruence (congruent, incongruent) × Electrode. Omnibus ANOVAs with 21 electrodes were used in each window, followed by ANOVAs including 16 electrodes for

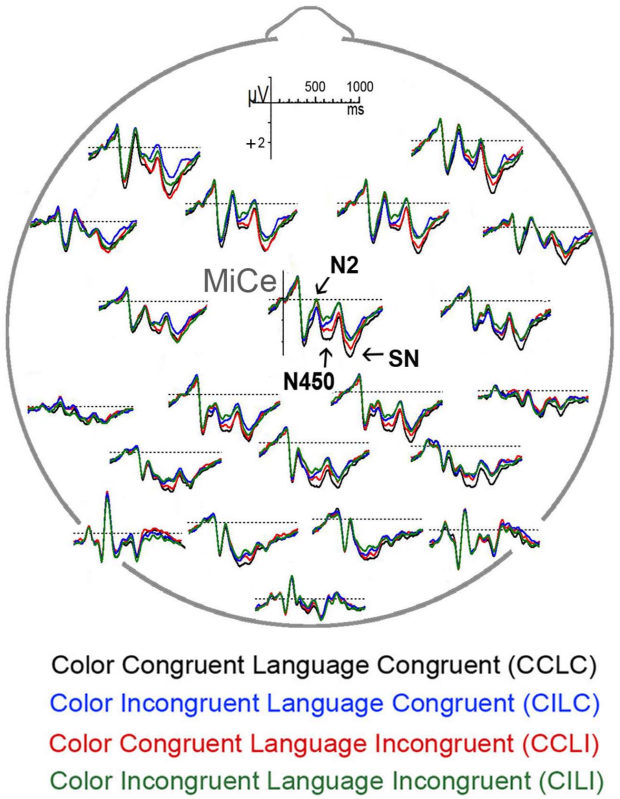


FIGURE 5 | A bird's eye view of the geodesic electrode array showing grand average ERPs. Voltage is plotted in microvolts on the *y*-axis with negative up; time is plotted in milliseconds on the *x*-axis with 0 ms marking the onset of the visually presented words and 100 ms tick marks.

scalp distribution analyses, with factors of Hemisphere (left, right), Anteriority (frontal, central, occipital), and Laterality (medial, lateral). In addition, region of interest analyses were used as appropriate for each effect. Effects for repeated-measures with greater than one degree of freedom are reported after Greenhouse–Geisser correction; planned contrasts were Bonferroni adjusted for multiple comparisons.

N2 (200–350 ms). **Figure 6** shows grand average ERPs at representative electrodes and a spline-interpolated scalp topography for the effect of language congruence. The omnibus ANOVA revealed a trend toward an effect of Language Congruence [$F(1, 23) = 3.625$; $p = 0.070$; Language Congruence by Electrode, $F(20, 460) = 2.214$; $p = 0.062$].

The distributional analysis revealed a Language Congruence by Laterality interaction [$F(1, 23) = 4.521$; $p = 0.044$] with a larger negativity for language incongruent than congruent trials that was significant at medial sites ($p = 0.039$) in planned contrasts. In *post hoc* analyses, data from medio-central and right-dorsal electrodes, which encompass the N2 distribution (LMFr, LMCe, RMFr, RMCe, RDFr, RDCe, MiCe, MiPa), were subjected to repeated-measures ANOVA. This confirmed that language incongruent trials elicited more negative amplitude than congruent trials over this region [Language Congruence, $F(1, 23) = 5.820$, $p = 0.024$].

N450 (350–550 ms). As expected, the omnibus ANOVA revealed a color-Stroop effect with a larger negativity for color incongruent than congruent trials, see **Figure 6** [Color Congruence, $F(1, 23) = 5.120$, $p = 0.033$; Color Congruence by Electrode, $F(20, 460) = 4.744$, $p = 0.001$]. Distributional analyses revealed the color-Stroop effect was present only at medial sites ($p = 0.003$) across all levels of anteriority with the strongest effect at medial central sites (Frontal, $p = 0.006$; Central $p = 0.002$; Occipital, $p = 0.013$), [Color Congruence × Laterality, $F(1, 23) = 15.806$, $p < 0.001$, Color Congruence × Laterality × Anteriority, $F(2, 46) = 3.384$, $p = 0.055$].

Sustained negativity (550–700 ms). The omnibus ANOVA revealed a color-Stroop effect with larger negativity for color incongruent than color congruent trials [Color Congruence, $F(1, 23) = 8.058$, $p = 0.009$], and a significant interaction between Color Congruence and Electrode [$F(20, 460) = 4.118$, $p = 0.014$], **Figure 6**[7]. The distributional analysis yielded a Color Congruence by Laterality interaction that showed the effect to be present at medial, but not lateral recording sites [$F(1, 23) = 6.927$, $p = 0.015$], and a Color Congruence by Anteriority interaction which revealed an effect at Frontal and Central, but not Occipital sites [$F(2, 46) = 5.017$, $p = 0.032$].

The interaction between Color Congruence and Language Congruence trended toward significance, $F(1, 23) = 3.717$, $p = 0.066$. **Figure 7** shows what appears to be an increased negativity as early as 400 ms for the within language Stroop effect compared to the between language Stroop effect. A sliding window analysis in 50 ms increments across the head revealed that both the between and within language Stroop effects were significant from 550 to 600 ms post-stimulus. Then the between language effect disappeared between 600 and 650 ms leading to a brief interaction between Color Congruence and Language Congruence [$F(1, 23) = 5.046$, $p = 0.035$], while the negativity for the Color Congruence effect within language continued through 700 ms.

Discussion

The goal of Experiment 2 was to study the temporal dynamics, and the corresponding neural and cognitive correlates, of the bilingual Stroop. The findings have implications for explaining the Stroop effect, both for bilinguals and monolinguals. Our data speak to the suggestion that the bilingual Stroop effect reflects a response set effect. We discuss the implications of our findings after a brief summary.

A large N450 effect was observed for the color congruence manipulation, replicating monolingual findings. Color incongruent trials elicited larger negative amplitude than color congruent trials between 350 and 550 ms post-stimulus onset. This effect was the same within and between languages, indicating that the N450 was sensitive to color congruence regardless of whether the distracters were from the response set or not. Following the N450, there was an effect of color congruence with SN amplitude for color incongruent compared to congruent trials. This

[7]Complex interactions with Naming Language in the distribution analysis could be explained by the loss of trials in Spanish (see Methods) and were not analyzed further.

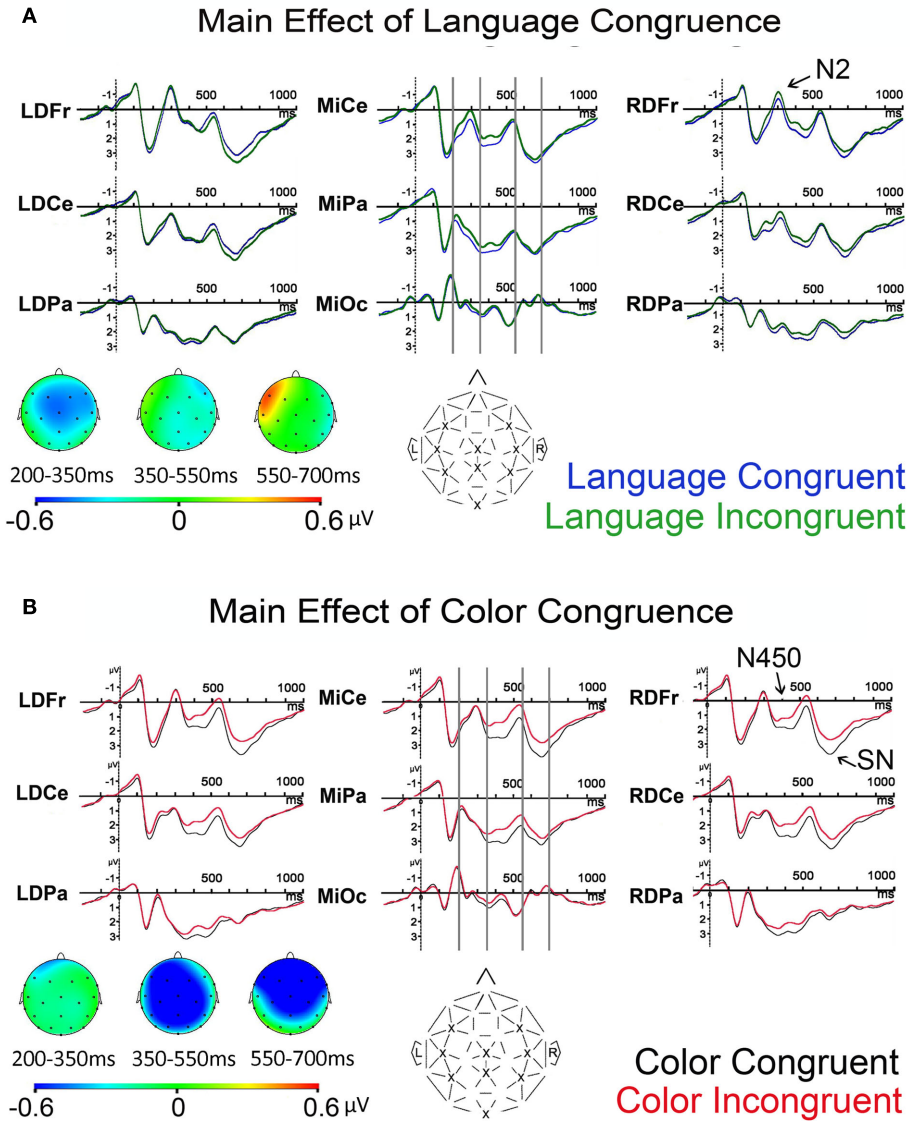


FIGURE 6 | Grand average ERPs for nine representative recording sites and spline-interpolated scalp topographies showing of three measured time windows for language congruence in (A) and color congruence in (B) (note that projection toward prefrontal channels is estimate). Vertical gray lines mark the time windows used for analyses for the N2, N450, and sustained negativity (SN). Electrode labeled from left to right: left frontal, central, parietal, medial central (vertex), parietal and occipital, right frontal, central, parietal.

effect was observed in the same time window as the conflict SP (550–700 ms post-stimulus onset), but did not share the typical distribution reported in monolingual studies (a sustained positivity over central–parietal scalp sites that reverses in polarity over lateral frontal sites; West, 2003; Markela-Lerenc et al., 2004). Finally, there was a language congruence effect at the N2 (200–350 ms), with greater negativity for between than within language trials. This effect was present at central and right frontal sites. The N2 was not modulated by color congruence.

A majority of monolingual Stroop ERP studies suggests that the N450 reflects response conflict and the SP reflects both response and stimulus level conflict. In particular, based on Chen et al. (2011), response-irrelevant items, such as between language distracters, should elicit response conflict, and any form of conflict should elicit effects in the subsequent time window. We found the opposite pattern of effects. The N450 was not significantly modulated by language congruence, with a strong effect for both between- and within language naming, while the SN was. If indeed the N450 reflects cognitive control related to response conflict, then our data indicate that color incongruent words created equal conflict and cognitive control demands regardless of whether they belonged to the response set or not. This is not to say that the N450 is completely insensitive to language congruence, or perhaps even to response set effects more generally. In fact, there appear to be hints of an interaction between color and language congruence, for example at vertex (MiCe) in **Figures 3** and **5**, although the

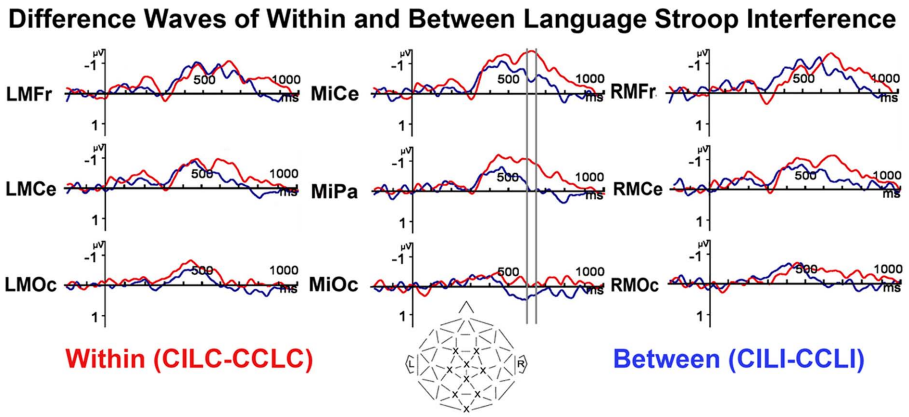


FIGURE 7 | Difference ERPs (color incongruence minus congruent) for within and between language trials separately. Sliding window analysis in 50 ms increments revealed that through 600 ms both between and within language Stroop effects were significant and no different from each other, then from 600 to 650 ms (highlighted with gray vertical bars) there was a brief interaction between color and language congruence, where only the within language Stroop effect was present.

interaction did not even approach significance at these locations (with *p*-values of 0.5–0.8 across the time window). Perhaps balanced bilinguals present a unique case in which the cross language lexical equivalents for the response set create response conflict at the N450. A critical test of this in future research would be to include words in both languages in line with the typical response set effects (e.g., PINK/*ROSA*), so that the degree of spread of activation between words within and between languages could be measured. Likewise, perhaps unbalanced bilinguals might show an N450 asymmetry across languages, with a larger effect for reading response relevant items in the dominant than non-dominant language – a testable question for future research.

Another characteristic of the N450 in this balanced bilingual sample is the broader distribution compared to monolinguals, which might reflect recruitment of additional neural substrates to process the dueling sources of interference (color and language) in the bilingual paradigm. There is growing evidence that bilinguals activate information in both of their languages even when using only one (Marian and Spivey, 2003; Kroll et al., 2006; Sunderman and Kroll, 2006; Duyck et al., 2007; Thierry and Wu, 2007). Consequently, to produce a word in the target language, bilinguals must inhibit the competing non-target language (Green, 1986, 1998; Meuter and Allport, 1999; Bialystok and Martin, 2004; Costa et al., 2006; Kroll et al., 2008; Hernández et al., 2010). Due to this demand, bilinguals may develop an inhibitory control mechanism that is specialized for language (Green, 1998) or domain-general (Roelofs et al., 2011) with benefits for inhibitory control on a variety of non-linguistic tasks, such as the Stroop, Simon, and card sorting tasks (Bialystok and Martin, 2004; Bialystok et al., 2004; Bialystok and Craik, 2010). Costa and Santesteban (2004) have suggested that benefits to executive control are moderated by proficiency across languages; while unbalanced bilinguals rely on inhibitory control to limit access, balanced bilinguals use a language-specific selection mechanism to control cross language interference. This suggestion is perhaps in line with Stroop performance in monolinguals, for whom a steady increase in the amount of Stroop interference is observed until attaining a third grade reading level (Comalli et al., 1962; Schiller, 1966), after which greater reading skill decreases the magnitude of the Stroop effect (Protopapas et al., 2007), reflecting gains in executive function and attentional control (Tzelgov et al., 1990). However, the between language N450 effect found in the current study suggests that the non-target language continues to be processed (beyond the N2), even on a task that does not require more than word form processing (c.f., Rodriguez-Fornells et al., 2002), and even for a response set that has minimal cross language orthographic overlap. Hence, the presence of an N450 Stroop effect both between and within languages lends support for non-selective activation of both languages in balanced bilinguals.

The results also reveal that language membership information is processed prior to the N450 – specifically at the N2. The N2 is thought to be a complex of components that are functionally and distributionally distinct based on stimuli and task demands (for a review of N2 findings, see Folstein and Van Petten, 2008). Most relevant for the current study, the N2 has sometimes been associated with early processes at the level of word form (see also Grainger et al., 2006, for a related component for word recognition). Larger N2 amplitude has been observed to word form information when attended than when not attended (Ruz and Nobre, 2008). By inference then, the attended response relevant language in the current study should have elicited larger N2 amplitude than the response-irrelevant language. We observed the opposite effect, indicating that the N2 observed herein is not related to attention to the response set (c.f. Lamers et al., 2010). Another possibility is that the N2 reflects conflict detection, such as that observed on the Erikson flanker task where both stimulus and response level conflict have resulted in an increase in N2 amplitude (Van Veen and Carter, 2002; Carter and Van Veen, 2007; Wendt et al., 2007). Our data are again inconsistent with the direction of this modulation, since within language trials create more conflict in the behavioral results, and by inference should elicit larger N2 amplitude.

Instead, our data is most consistent with a third type of N2 effect. The direction and scalp distribution of the N2 effect in the current study (slight right-lateralization with a fronto-central

maximum; c.f. Aron et al., 2003) is more in line with a no-go N2 (Pliszka et al., 2000, 2007; Liotti et al., 2007), than with either an attentional set effect or a conflict N2. The no-go N2 typically shows larger negative amplitude related to inhibiting a response (Pliszka et al., 2000; Schmajuk et al., 2006; Woodward et al., 2007; Folstein and Van Petten, 2008). In the bilingual Stroop paradigm, within language items are all potential go candidates as part of the response set, while between language distracters are all no-go items. Thus language membership is recognized early, presumably based on word form information, triggering mechanisms of inhibition as reflected by a no-go N2 for between language distracters. Yet, inhibition of the response for between language trials cannot completely explain our data. First, response relevant distracters should also elicit a no-go N2 relative to congruent trials. Our design does not have the power to determine if there is a no-go effect for within language distracters, but future research may show a graded effect for inhibition of response relevant and irrelevant items across languages. Second, clearly this stage of processing does not reflect complete inhibition of between language distracters given the subsequent N450 and SN. Instead, it may reflect a stage of processing parallel to that of the N450, which together may contribute to the end-state behavioral bilingual Stroop effect.

The behavioral findings from Experiments 1 and 2 were consistent with the majority of the literature, showing a larger color word Stroop effect within language than between languages (MacLeod, 1991; Francis, 1999). For this reason, the most surprising effect, or lack thereof, in Experiment 2 was the absence of a clear interaction between color and language congruence. If not from a direct interaction at the N450 or earlier brain activity, where does the interaction between color and language in the RTs come from? It is possible that ERP technology is not sensitive to the source of the BWLS, if for example it is driven by weak or deep sources of brain activity (or sources that cancel at the scalp). This seems unlikely given that our data show robust effects for both color and language congruence that are inline with previous findings. Instead, our data seem to indicate that color and language conflict are processed independently at different time intervals and interact only for a fleeting moment during the late time window of the SN.

It is possible that the BWLS is purely due to the underlying processes reflected in the brief interaction at the late SN. Our data reflect a broadly distributed, SN, inline with earlier reports of ERP effects in a complex Stroop task (West and Alain, 1999). Despite the similarity in scalp distribution, it is unlikely that the SN is simply sustained activity from the N450. The SN appears to resolve more quickly between than within languages. Perhaps this negativity is functionally related to the conflict SP, thought to reflect response monitoring and conflict adaptation (West and Alain, 2000; West, 2003; West et al., 2005; Chen et al., 2011). It could result from a global difference trial to trial in conflict adaptation, with quicker adaptation to between than within language conflict, or a greater impact of response relevant words on response monitoring. Still, these processes must be triggered by earlier stages of processing in which detection occurs of the conflict within or between languages. Perhaps this earlier stage of processing is reflected in the N2 effect. Thus, rather than complete inhibition of the between language distracters, the N2 may index processes of inhibitory control that facilitate later resolution of conflict at the SN. Between language distracters trigger this early inhibitory (no-go) mechanism, resulting in a larger N2 and subsequently quicker resolution of the SN. The intermediate effect at the N450 must then reflect parallel processing of the distracter words, regardless of response set (or language) membership. Thus, the behavioral bilingual Stroop effect could be a product of activity across parallel processing of language and color rather than the presence of a direct interaction of the two. In other words, it is possible that the RT effects reflect the summed brain activity over time, with contributions from language conflict and color conflict at different points in time (c.f., Cohen et al., 1990; Roelofs, 2003).

CONCLUSION

In summary, data from two bilingual Stroop experiments aimed at uncovering the source of the well-documented bilingual Stroop effect – referred to herein as the between-within language Stroop effect or BWLS. Experiment 1 replicated the BWLS in both balanced and unbalanced bilinguals. This effect was present regardless of language dominance, and during both single language and mixed language presentation. However, by taking an unconventional look at the Stroop data, analyzing the effect of language congruence in the presence or absence of color-Stroop interference, we were able to show that the source of the BWLS varied based on these manipulations. In the process of thoroughly delineating the behavior of our population on the bilingual Stroop task, we were able to address the leading explanation for the BWLS. We show that a response set effect can only partially explain this effect. Experiment 2 delineated the time course and stage of processing at which the BWLS occurs using a real time electrophysiological measure. Our ERP data provide evidence that balanced bilinguals process language congruence prior to color congruence on a bilingual color word Stroop task, as indexed by a language effect at the N2. Importantly, distinguishing the distracters based on language did not affect later processes at the N450, indicating that color incongruent words created equal conflict and cognitive control demands regardless of whether they belonged to the response set or not. Rather than complete inhibition of the between language distracters, the N2 may reflect processes of inhibitory control that facilitate the resolution of conflict at the SN, while the N450 reflects parallel processing of the distracter words, regardless of response set (or language). In sum, the behavioral BWLS reflects summed brain activity over time, with contributions from language conflict and color conflict at different time points. Our findings add to a vast literature, informing models of both monolingual and bilingual conflict processing on the Stroop task, and present new questions for the field.

ACKNOWLEDGMENTS

We have many individuals to thank for advice and technical assistance on this project, including Ryan J. Giuliano, Amanda Martinez-Lincoln, Shukhan Ng, David Pillow, Elena Salillas, Mai-Anh Tran Ngoc, and especially Delia Kothmann Paskos who inspired this research study. Resources and support were provided by the Computational Biology Initiative. Funding was provided by NICHD/NIGMH HD060435 and the UTSA College of Science to N. Y. Y. Wicha.

REFERENCES

Aarts, E., Roelofs, A., and van Turennout, M. (2009). Attentional control of task and response in lateral and medial frontal cortex: brain activity and reaction time distributions. *Neuropsychologia* 47, 2089–2099.

Abutalebi, J., Annoni, J. M., Zimine, I., Pegna, A. J., Seghier, M. L., Lee-Jahnke, H., Lazeyras, F., Cappa, S. F., and Khateb, A. (2008). Language control and lexical competition in bilinguals: an event-related fMRI study. *Cereb. Cortex* 18, 1496–1505.

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., and Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nat. Neurosci.* 6, 115–116.

Atkinson, C. M., Drysdale, K. A., and Fulham, W. R. (2003). Event-related potentials to Stroop and reverse Stroop stimuli. *Int. J. Psychophysiol.* 47, 1–21.

Bialystok, E., Craik, F. I., Klein, R., and Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychol. Aging* 19, 290–303.

Bialystok, E., and Craik, F. I. M. (2010). Cognitive and linguistic processing in the bilingual mind. *Curr. Dir. Psychol. Sci.* 19, 19–23.

Bialystok, E., Craik, F. I. M., and Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 859–873.

Bialystok, E., and Martin, M. M. (2004). Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. *Dev. Sci.* 7, 325–339.

Carter, C. S., and Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cogn. Affect. Behav. Neurosci.* 7, 367–379.

Chen, A., Bailey, K., Tiernan, B. N., and West, R. (2011). Neural correlates of stimulus and response interference in a 2-1 mapping Stroop task. *Int. J. Psychophysiol.* 80, 129–138.

Chen, H., and Ho, C. (1986). Development of Stroop Interference in Chinese-English Bilinguals. *J. Exp. Psychol. Learn. Mem. Cogn.* 12, 397–401.

Christoffels, I. K., Firk, C., and Schiller, N. O. (2007). Bilingual language control: an event-related brain potential study. *Brain Res.* 1147, 192–208.

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol. Rev.* 97, 332–361.

Comalli, P. E. Jr., Wapner, S., and Werner, H. (1962). Interference effects of Stroop color-word test in childhood, adulthood, and aging. *J. Gen. Psychol.* 100, 47–53.

Costa, A., and Santesteban, M. (2004). Lexical access in bilingual speech production: evidence from language switching in highly proficient bilinguals and L2 learners. *J. Mem. Lang.* 50, 491–451.

Costa, A., Santesteban, M., and Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1057–1074.

Dalrymple-Alford, E. C. (1968). Interlingual interference in a color-naming task. *Psychon. Sci.* 10, 215–216.

Dijkstra, T., and Van Heuven, W. J. (2002). Modeling bilingual word recognition: past, present and future: reply. *Biling. (Camb. Engl.)* 5, 219–224.

Dijkstra, T., Van Heuven, W. J., and Grainger, J. (1998). Simulating cross-language competition with the bilingual interactive activation model. *Psychol. Belg.* 38, 177–196.

Duncan-Johnson, C. C., and Kopell, B. S. (1981). The Stroop effect – brain potentials localize the source of interference. *Science* 214, 938–940.

Duyck, W., Van Assche, E., Drieghe, D., and Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: evidence for nonselective access. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 663–679.

Dyer, F. N. (1971). Color-naming interference in monolinguals and bilinguals. *J. Verb. Learn. Verb. Behav.* 10, 297–302.

Folstein, J. R., and Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology* 45, 152–170.

Francis, W. S. (1999). Cognitive integration of language and memory in bilinguals: semantic representation. *Psychol. Bull.* 125, 193–222.

Gasquoine, P. G., Croyle, K. L., Cavazos-Gonzalez, C., and Sandoval, O. (2007). Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults. *Arch. Clin. Neuropsychol.* 22, 991–1001.

Glaser, W. R., and Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *J. Exp. Psychol. Gen.* 118, 13–42.

Goldfarb, L., and Tzelgov, J. (2007). The cause of the within-language Stroop superiority effect and its implications. *Q. J. Exp. Psychol.* 60, 179–185.

Grainger, J., Kiyonaga, K., and Holcomb, P. J. (2006). The time course of orthographic and phonological code activation. *Psychol. Sci.* 17, 1021–1026.

Green, D. W. (1986). Control, activation and resource: a framework and a model for the control of speech in bilinguals. *Brain Lang.* 27, 210–223.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Biling. (Camb. Engl.)* 1, 67–81.

Green, D. W., Grogan, A., Crinion, J., Ali, N., Sutton, C., and Price, C. J. (2010). Language control and parallel recovery of language in individuals with aphasia. *Aphasiology* 24, 188–209.

Grosjean, F. (1998). Studying bilinguals: methodological and conceptual issues. *Biling. (Camb. Engl.)* 1, 131–149.

Hennighausen, E., Heil, M., and Rosler, F. (1993). A correction method for DC drift artifacts. *Electroencephalogr. Clin. Neurophysiol.* 86, 199–204.

Hernandez, A. E. (2009). Language switching in the bilingual brain: what's next? *Brain Lang.* 109, 133–140.

Hernández, M., Costa, A., Fuentes, L. J., Vivas, A. B., and Sebastián-Gallés, N. (2010). The impact of bilingualism on the executive control and orienting networks of attention. *Biling. (Camb. Engl.)* 13, 315–325.

Ilan, A. B., and Polich, J. (1999). P300 and response time from a manual Stroop task. *Clin. Neurophysiol.* 110, 367–373.

Jackson, G. M., Swainson, R., Cunnington, R., and Jackson, S. R. (2001). ERP correlates of executive control during repeated language switching. *Biling. Lang. Cognit.* 4, 169–178.

Jared, D., and Kroll, J. F. (2001). Do bilinguals activate phonological representations in one or both of their languages when naming words? *J. Mem. Lang.* 44, 2–31.

Jolles, J., Houx, P. J., van Boxtel, M. P. J., and Ponds, R. W. H. M. (1995). *Maastricht Aging Study: Determinants of Cognitive Aging.* Maastricht: Neuropsych Publishers.

Kaplan, H., Goodglass, S., and Weintraub, E. (2001). *The Boston Naming Test.* Philadelphia, PA: Lippincott Williams, and Wilkins.

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *Am. J. Psychol.* 77, 576–588.

Kroll, J. F., Bobb, S., and Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: arguments against a fixed locus of language selection in bilingual speech. *Biling. (Camb. Engl.)* 9, 119–135.

Kroll, J. F., Bobb, S. C., Misra, M., and Guo, T. (2008). Language selection in bilingual speech: evidence for inhibitory processes. *Acta Psychol. (Amst.)* 128, 416–430.

Kroll, J. F., and Stewart, E. (1994). Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *J. Mem. Lang.* 33, 149–174.

Kroll, J. F., Van Hell, J. G., Tokowicz, N., and Green, D. W. (2010). The revised hierarchical model: a critical review and assessment. *Biling. (Camb. Engl.)* 13, 373–381.

Kutas, M., Mccarthy, G., and Donchin, E. (1977). Augmenting mental chronometry – P300 as ameasure of stimulus evaluation time. *Science* 197, 792–795.

Lamers, M., Roelofs, A., and Rabeling-Keus, I. (2010). Selective attention and response set in the Stroop task. *Mem. Cognit.* 38, 893–904.

Lansbergen, M. M., and Kenemans, J. L. (2008). Stroop interference and the timing of selective response activation. *Clin. Neurophysiol.* 119, 2247–2254.

Larson, M. J., Kaufman, D. A. S., and Perlstein, W. M. (2009). Neural time course of conflict adaptation effects on the Stroop task. *Neuropsychologia* 47, 663–670.

Lemhöfer, K., and Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: effects of code similarity in language-specific and generalized lexical decision. *Mem. Cognit.* 32, 533–550.

Liotti, M., Pliszka, S. R., Perez, R., Luus, B., Glahn, D., and Semrud-Clikeman, M. (2007). Electrophysiological correlates of response inhibition in children and adolescents with ADHD: influence of gender, age, and previous treatment history. *Psychophysiology* 44, 936–948.

Liotti, M., Woldorff, M. G., Perez, R., and Mayberg, H. S. (2000). An ERP study of the temporal course of the Stroop color-word interference effect. *Neuropsychologia* 38, 701–711.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychol. Bull.* 109, 163–203.

MacLeod, C. M., and MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention. *Trends Cogn. Sci. (Regul. Ed.)* 4, 383–391.

Mägiste, E. (1984). Stroop tasks and dichotic translation: the development of interference patterns in bilinguals. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 304–315.

Mägiste, E. (1985). Development of intralingual and interlingual interference in bilinguals. *J. Psycholinguist. Res.* 14, 137–154.

Marian, V., and Spivey, M. (2003). Competing activation in bilingual language processing: within- and between-language competition. *Biling. (Camb. Engl.)* 6, 97–115.

Markela-Lerenc, J., Ille, N., Kaiser, S., Fiedler, P., Mundt, C., and Weisbrod, M. (2004). Prefrontal-cingulate activation during executive control: which comes first? *Brain Res. Cogn. Brain Res.* 18, 278–287.

Meuter, R. F. I., and Allport, A. (1999). Bilingual language switching in naming: asymmetrical costs of language selection. *J. Mem. Lang.* 40, 25–40.

Midgley, K. J., Holcomb, P. J., and Grainger, J. (2009). Language effects in second language learners and proficient bilinguals investigated with event-related potentials. *J. Neurolinguistics* 22, 281–300.

Moreno, E. M., Federmeier, K. D. M., and Kutas, M. (2002). Switching languages, switching palabras (words): an electrophysiological study of code switching. *Brain Lang.* 80, 188–207.

Okuniewska, H. (2007). Impact of second language proficiency on the bilingual Polish-English Stroop task. *Psychol. Lang. Commun.* 11, 49–63.

Peckham, A. D., McHugh, R. K., and Otto, M. W. (2010). A meta-analysis of the magnitude of biased attention in depression. *Depress. Anxiety* 27, 1135–1142.

Pliszka, S. R., Liotti, M., Bailey, B. Y., Perez, R., Glahn, D. C., and Semrud-Clikeman, M. (2007). Electrophysicological effects of stimulant treatment on inhibitory control in children with attention-deficit/hyperactivity disorder. *J. Child Adolesc. Psychopharmacol.* 17, 356–366.

Pliszka, S. R., Liotti, M., and Woldorff, M. G. (2000). Inhibitory control in children with attention-deficit/hyperactivity disorder: event-related potentials identify the processing component and timing of an impaired right-frontal response-inhibition mechanism. *Biol. Psychiatry* 48, 238–246.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148.

Preston, M. S., and Lambert, W. E. (1969). Interlingual interference in a bilingual version of the Stroop color-word task. *J. Verb. Learn. Verb. Behav.* 8, 295–301.

Proctor, R. W. (1978). Sources of color-word interference in the Stroop color-naming task. *Percept. Psychophys.* 23, 413–419.

Protopapas, A., Archonti, A., and Skaloumbakas, C. (2007). Reading ability is negatively related to Stroop interference. *Cogn. Psychol.* 54, 251–282.

Proverbio, A. M., Adorni, R., and Zani, A. (2009). Inferringnative language from early bio-electricalactivity. *Biol. Psychol.* 80, 52–63.

Pukrop, R., and Klosterkötter, J. (2010). Neurocognitive indicators of clinical high-risk states for psychosis: a critical review of the evidence. *Neurotox. Res.* 18, 272–286.

Rebai, M., Bernard, C., and Lannou, J. (1997). The Stroop's test evokes a negative brain potential, the N400. *Int. J. Neurosci.* 91, 85–94.

Rodriguez-Fornells, A., Rotte, M., Heinze, H. J., Nosselt, T., and Munte, T. F. (2002). Brain potential and functional MRI evidence for how to handle two languages with one brain. *Nature* 415, 1026–1029.

Rodriguez-Fornells, A., Van Der Lugt, A., Rotte, M., Britti, B., Heinze, H. J., and Munte, T. F. (2005). Second language interferes with word production in fluent bilinguals: brain potential and functional imaging evidence. *J. Cogn. Neurosci.* 17, 422–433.

Roelofs, A. (2003). Goal-referenced selection of verbal action: modeling attentional control in the Stroop task. *Psychol. Rev.* 110, 88–125.

Roelofs, A. (2010). Attention and facilitation: converging information versus inadvertent reading in Stroop task performance. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 411–422.

Roelofs, A., Piai, V., and Garrido Rodriguez, G. (2011). Attentional inhibition in bilingual naming performance: evidence from delta-plot analyses. *Front. Psychol.* 2:184. doi:10.3389/fpsyg.2011.00184

Rosenfeld, J. P., and Skogsberg, K. R. (2006). P300-based Stroop study with low probability and target Stroop oddballs: the evidence still favors the response selection hypothesis. *Int. J. Psychophysiol.* 60, 240–250.

Rosselli, M., Ardila, A., Santisi, M. N., Arecco Mdel, R., Salvatierra, J., Conde, A., and Lenis, B. (2002). Stroop effect in Spanish-English bilinguals. *J. Int. Neuropsychol. Soc.* 8, 819–827.

Ruz, M., and Nobre, A. C. (2008). Attention modulates initial stages of visual word processing. *J. Cogn. Neurosci.* 20, 1727–1736.

Schiller, P. H. (1966). Developmental study of color-word interference. *J. Exp. Psychol.* 72, 105–108.

Schmajuk, M., Liotti, M., Busse, L., and Woldorff, M. G. (2006). Electrophysiological activity underlying inhibitory control processes in normal adults. *Neuropsychologia* 44, 384–395.

Spivey, M. J., and Marian, V. (1999). Cross talk between native and second languages: partial activation of an irrelevant lexicon. *Psychol. Sci.* 10, 281–284.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol.* 18, 643–662.

Sumiya, H., and Healy, A. F. (2004). Phonology in the bilingual Stroop effect. *Mem. Cognit.* 32, 752–758.

Sunderman, G., and Kroll, J. F. (2006). First language activation during second language lexical processing: an investigation of lexical form, meaning, and grammatical class. *Stud. Sec. Lang. Acquis.* 28, 387–422.

Thierry, G., and Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12530–12535.

Treisman, A. M., and Fearnley, S. (1969). The Stroop test: selective attention to colours and words. *Nature* 222, 437–439.

Tzelgov, J., Henik, A., and Leiser, D. (1990). Controlling Stroop interference: evidence from a bilingual task. *J. Exp. Psychol. Learn. Mem. Cogn.* 16, 760–771.

Van Veen, V., and Carter, C. S. (2002). The anterior cingulate as a conflict monitor: fMRI and ERP studies. *Physiol. Behav.* 77, 477–482.

van Veen, V., and Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: a functional MRI study. *Neuroimage* 27, 497–504.

Wendt, M., Heldmann, M., Munte, T. F., and Kluwe, R. H. (2007). Disentangling sequential effects of stimulus- and response-related conflict and stimulus-response repetition using brain potentials. *J. Cogn. Neurosci.* 19, 1104–1112.

West, R. (2003). Neural correlates of cognitive control and conflict detection in the Stroop and digit-location tasks. *Neuropsychologia* 41, 1122–1135.

West, R., and Alain, C. (1999). Event-related neural activity associated with the Stroop task. *Brain Res. Cogn. Brain Res.* 8, 157–164.

West, R., and Alain, C. (2000). Effects of task context and fluctuations of attention on neural activity supporting performance of the Stroop task. *Brain Res.* 873, 102–111.

West, R., Jakubek, K., Wymbs, N., Perry, M., and Moore, K. (2005). Neural correlates of conflict processing. *Exp. Brain Res.* 167, 38–48.

West, R., Krompinger, J., Bowry, R., and Doll, R. (2004). Neural correlates of conflict monitoring and error processing. *Brain Cogn.* 54, 168–170.

Woodward, T. S., Buchy, L., Moritz, S., and Liotti, M. (2007). A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. *Schizophr. Bull.* 33, 1023–1028.

Zied, K. M., Phillipe, A., Karine, P., Valerie, H. T., Ghislaine, A., Arnaud, R., and Didier, L. G. (2004). Bilingualism and adult differences in inhibitory mechanisms: evidence from a bilingual Stroop task. *Brain Cogn.* 54, 254–256.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 September 2011; accepted: 02 March 2012; published online: 02 April 2012.

*Citation: Naylor LJ, Stanley EM and Wicha NYY (2012) Cognitive and electrophysiological correlates of the bilingual Stroop effect. Front. Psychology **3**:81. doi: 10.3389/fpsyg.2012.00081*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 February 2012
doi: 10.3389/fpsyg.2012.00039

Effects of speech rate and practice on the allocation of visual attention in multiple object naming

Antje S. Meyer[1]*, Linda Wheeldon[2], Femke van der Meulen[2] and Agnieszka Konopka[3]

[1] *Max Planck Institute for Psycholinguistics and Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, Netherlands*
[2] *School of Psychology, University of Birmingham, Birmingham, UK*
[3] *Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Robert J. Hartsuiker, University of Ghent, Belgium
Zenzi M. Griffin, University of Texas, USA

***Correspondence:**
Antje S. Meyer, Max Planck Institute for Psycholinguistics and Donders Institute for Brain, Cognition and Behavior, Radboud University, Postbus 310, 6500 AH Nijmegen, Netherlands.
e-mail: antje.meyer@mpi.nl

Earlier studies had shown that speakers naming several objects typically look at each object until they have retrieved the phonological form of its name and therefore look longer at objects with long names than at objects with shorter names. We examined whether this tight eye-to-speech coordination was maintained at different speech rates and after increasing amounts of practice. Participants named the same set of objects with monosyllabic or disyllabic names on up to 20 successive trials. In Experiment 1, they spoke as fast as they could, whereas in Experiment 2 they had to maintain a fixed moderate or faster speech rate. In both experiments, the durations of the gazes to the objects decreased with increasing speech rate, indicating that at higher speech rates, the speakers spent less time planning the object names. The eye–speech lag (the time interval between the shift of gaze away from an object and the onset of its name) was independent of the speech rate but became shorter with increasing practice. Consistent word length effects on the durations of the gazes to the objects and the eye-speech lags were only found in Experiment 2. The results indicate that shifts of eye gaze are often linked to the completion of phonological encoding, but that speakers can deviate from this default coordination of eye gaze and speech, for instance when the descriptive task is easy and they aim to speak fast.

Keywords: utterance planning, speech rate, practice, eye movements, visual attention

INTRODUCTION

We can talk in different ways. We can, for instance, use a special register, child directed speech, to talk to a child, and we tend to deliver speeches, and formal lectures in a style that is different from casual dinner table conversations. The psychological processes underlying the implementation of different speech styles have rarely been studied. The present paper concerns one important feature distinguishing different speech styles, i.e., speech rate. It is evident that speakers can control their speech rate, yet little is known about how they do this.

To begin to explore this issue we used a simple speech production task: speakers named sets of pictures in sequences of nouns (e.g., "kite, doll, tap, sock, whale, globe"). Each set was shown on several successive trials. In the first experiment, the speakers were asked to name the pictures as fast as they could. In the second experiment, they had to maintain a fixed moderate or faster speech rate, which allowed us to separate the effects of speech rate and practice. Throughout the experiments, the speakers' eye movements were recorded along with their spoken utterances. In the next sections, we motivate this approach, discuss related studies, and explain the predictions for the experiments.

SPEECH-TO-GAZE ALIGNMENT IN DESCRIPTIVE UTTERANCES

In many language production studies participants have been asked to name or describe pictures of one or more objects. Though probably not the most common way of using language, picture naming is popular in language production research because it offers good control of the content of the speakers' utterances and captures a central component of speech planning, namely the retrieval of words from the mental lexicon.

In some picture naming studies, the speakers' eye movements were recorded along with their speech. This is useful because a person's eye gaze reveals where their visual attention is focused, that is, which part of the environment they are processing with priority (e.g., Deubel and Schneider, 1996; Irwin, 2004; Eimer et al., 2007). In picture naming, visual attention, and eye gaze are largely controlled endogenously (i.e., governed by the speaker's goals and intentions), rather than exogenously (i.e., by environmental stimuli). That is, speakers actively direct their gaze to the objects they wish to focus on. Therefore, eye movements provide not only information about the speaker's visual processing, but also, albeit more indirectly, about the executive control processes engaged in the task (for discussions of executive control processes see Baddeley, 1986; Posner and Petersen, 1990; Miyake et al., 2000).

The eye movement studies of language production have yielded a number of key findings. First, when speakers name sets of objects they typically look at each of the objects in the order of mention, just before naming it (e.g., Meyer et al., 1998; Griffin, 2001). When speakers describe cartoons of events or actions, rather than naming individual objects, there can be a brief apprehension phase during which speakers gain some understanding of the gist of the scene and during which their eye movements are not related in any obvious way to the structure of the upcoming utterances, but following this, there is again a tight coupling between eye gaze and

speech output, with each part of the display being inspected just before being mentioned (Griffin and Bock, 2000; Bock et al., 2003; but see Gleitman et al., 2007).

A second key result is that the time speakers spend looking at each object (hereafter, gaze duration) depends not only on the time they need for the visual–conceptual processing of the object (e.g., Griffin and Oppenheimer, 2006) but also on the time they require to select a suitable name for the object and to retrieve the corresponding word form. This has been shown in studies where the difficulty of identifying the objects, the difficulty of retrieving their names from the lexicon, or the difficulty of generating the corresponding word forms was systemically varied. All of these manipulations affected how long the participants looked at the objects (e.g., Meyer et al., 1998; Griffin, 2001, 2004; Belke and Meyer, 2007). For the present research, a particularly important finding is that speakers look longer at objects with long names than at objects with shorter names (e.g., Meyer et al., 2003, 2007; Korvorst et al., 2006; but see Griffin, 2003). This indicates that speakers usually only initiate the shift of gaze and attention to a new object after they have retrieved the name of the current object (Roelofs, 2007, 2008a,b). A likely reason for the late shifts of gaze and attention is that attending to an object facilitates not only its identification but also the retrieval of any associated information, including the object name (e.g., Wühr and Waszak, 2003; Wühr and Frings, 2008). This proposal fits in well with results demonstrating that lexical access is not an automatic process, but requires some processing capacity (e.g., Ferreira and Pashler, 2002; Cook and Meyer, 2008; Roelofs, 2008a,b) and would therefore benefit from the allocation of attention. The same should hold for speech-monitoring processes (for reviews and further discussion see Postma, 2000; Hartsuiker et al., 2005; Hartsuiker, 2006; Slevc and Ferreira, 2006), which are capacity demanding and might also benefit from focused visual attention to the objects being described (e.g., Oomen and Postma, 2002).

EMPIRICAL FINDINGS ON EYE–SPEECH COORDINATION AT DIFFERENT SPEECH RATES

The studies reviewed above demonstrated that during naming tasks, the speakers' eye movements are tightly coordinated in time with their speech planning processes, with speakers typically looking at each object until they have planned its name to the level of the phonological form. This coupling of eye gaze and speech planning is not dictated by properties of the visual or the linguistic processing system. Speakers can, of course, choose to coordinate their eye gaze and speech in different ways, moving their eyes from object to object sooner, for instance as soon as they have recognized the object, or much later, for instance after they have produced, rather than just planned, the object's name. In this section, we review studies examining whether the coordination of eye gaze and speech varies with speech rate. One would expect that when speakers aim to talk fast, they should spend less time planning each object name. Given that the planning times for object names have been shown to be reflected in the durations of the gazes to the objects, speakers should show shorter gaze durations at faster speech rates. In addition, the coordination of eye gaze and speech might also change. At higher speech rates, speakers might, for instance, plan further ahead, i.e., initiate the shift of gaze to a new object earlier relative to the onset of the object name, in order to insure the fluency of their utterances.

Spieler and Griffin (2006) asked young and older speakers (average ages: 20 vs. 75 years, respectively) to describe pictures in utterances such as "The crib and the limousine are above the needle." They found that the older speakers looked longer at the objects and took longer to initiate and complete their utterances than the younger ones. However, the temporal coordination of gaze with the articulation of the utterances was very similar for the two groups. Before speech onset, both groups looked primarily at the first object and spent similar short amounts of time looking at the second object. Belke and Meyer (2007, Experiment 1) obtained similar results. Older speakers spoke more slowly than younger speakers and inspected the pictures for longer, but the coordination between eye gaze and speech in the two groups was similar.

Mortensen et al. (2008) also found that older speakers spoke more slowly and looked at the objects for longer than younger speakers. However, in this study the older participants had shorter eye–speech lags than younger speakers. Griffin (2003) reported a similar pattern of results. She asked two groups of college students attending schools in different regions of the US to name object pairs in utterances such as "wig, carrot." For unknown reasons, one group of participants articulated the object names more slowly than the other group. Before speech onset, the slower talkers spent more time looking at the first object and less time looking at the second object than the fast talkers, paralleling the findings obtained by Mortensen and colleagues for older speakers. Thus, compared to the fast talkers, the slower talkers delayed the shift of gaze and carried out more of the phonetic and articulatory planning of the first object name while still attending to that object.

These studies involved comparisons of speakers differing in their habitual speech rates. By contrast, Belke and Meyer (2007, Experiment 2) asked one group of young participants to adopt a speech rate that was slightly higher than the average rate used by the young participants in an earlier experiment (Belke and Meyer, 2007, Experiment 1, see above) or a speech rate that was slightly lower than the rate adopted by older participants in that experiment. As expected, these instructions affected the speakers' speech rates and the durations of their gazes to the objects. In line with the results obtained by Mortensen et al. (2008) and by Griffin (2003), the eye–speech lag was much shorter at the slow than at the fast speech rate.

To sum up, in object naming tasks, faster speech rates are associated with shorter gazes to the objects. Given the strong evidence linking gaze durations to speech planning processes, these findings indicate that when speakers increase their speech rate, they spend less time planning their words (see also Dell et al., 1997). While some studies found no change in the coordination of eye gaze and speech, others found shorter eye–speech lags during slow than during faster speech. Thus, during slow speech, the shift of gaze from the current to the next object occurred later relative to the onset of current object name than during faster speech. It is not clear why this is the case. Perhaps slow speech is often carefully articulated speech and talkers delay the shift of gaze in order to carry out more of the phonetic and articulatory planning processes for an object name while still attending to that object. As Griffin

(2003) pointed out, speakers do not need to look ahead much in slow speech because they have ample time to plan upcoming words during the articulation of the preceding words.

THE PRESENT STUDY

Most of the studies reviewed above concerned comparisons between groups of speakers differing in their habitual speech rate. Interpreting their results is not straightforward because it is not known why the speakers preferred different speech rates. So far, the study by Belke and Meyer (2007) is, to our knowledge, the only one where eye movements were compared when one group of speakers used different speech rates, either a moderate or a very slow rate.

The goal of the present study was to obtain additional evidence about the way speakers coordinate their eye movements with their speech when they adopt different speech rates. Gaze durations indicate when and for how long speakers direct their visual attention to each of the objects they name. By examining the speaker's eye movements at different speech rates, we can determine how their planning strategies – the time spent planning each object name and the temporal coordination of planning and speaking – might change.

Whereas speakers in Belke and Meyer's (2007) study used a moderate or a very slow speech rate, speakers in the first experiment of present study were asked to increase their speech rate beyond their habitual rate and to talk as fast as they could. To the best of our knowledge no other study has used these instructions, though the need to talk fast regularly occurs in everyday conversations.

Participants saw sets of six objects each (see **Figure 1**) and named them as fast as possible. There were eight different sets, four featuring objects with monosyllabic names and four featuring objects with disyllabic names (see Appendix). In Experiment 1, there were two test blocks, in each of which each set was named on eight successive trials. We recorded the participants' eye movements and speech onset latencies and the durations of the spoken words. We asked the participants to name the same objects on successive trials (rather than presenting new objects on each trial) to make sure that they could substantially increase their speech rate without making too many errors. An obvious drawback of this procedure was that the effects of increasing speech rate and increasing familiarity with the materials on the speech-to-gaze coordination could not be separated. We addressed this issue in Experiment 2.

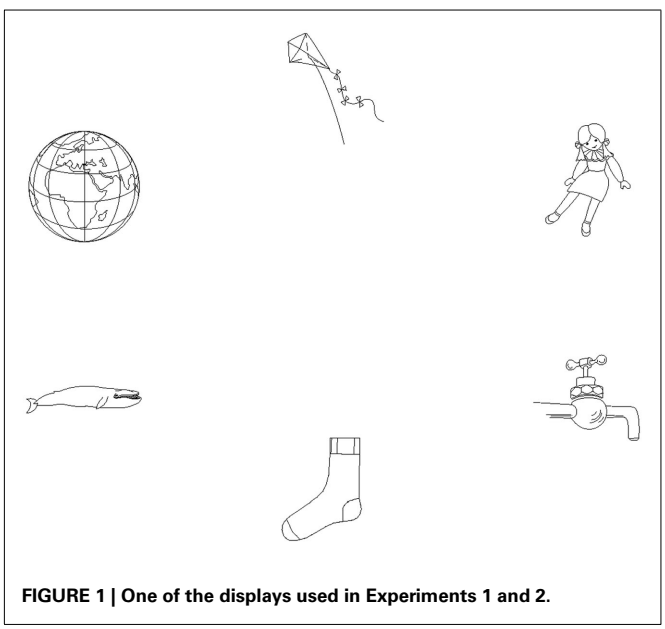


FIGURE 1 | One of the displays used in Experiments 1 and 2.

Based on the results summarized above, we expected that speakers would look at most of the objects before naming them and that the durations of the gazes to the objects would decrease with increasing speech rate. The eye–speech lags should either be unaffected by the speech rate or increase with increasing speech rate. That is, as the speech becomes faster speakers might shift their gaze earlier and carry out more of the planning of the current word without visual guidance.

We compared gaze durations for objects with monosyllabic and disyllabic names. As noted, several earlier eye tracking studies had shown that speakers looked longer at objects with long names than at objects with shorter names (e.g., Meyer et al., 2003, 2007; Korvorst et al., 2006; but see Griffin, 2003). This indicates that the speakers only initiated the shift of gaze to a new object after they had retrieved the phonological form of the name of the current object. In these studies no particular instructions regarding speech rate were given. If speakers consistently time the shifts of gaze to occur after phonological encoding of the current object name has been completed, the word length effect should be seen regardless of the speech rate. By contrast, if at high speech rates, speakers initiate the shifts of gaze from one object to the next earlier, before they have completed phonological encoding of the current object name, no word length effect on gaze durations should be seen.

EXPERIMENT 1

METHOD

Participants

The experiment was carried out with 24 undergraduate students of the University of Birmingham. They were native speakers of British English and had normal or corrected-to-normal vision. They received either payment or course credits for participation. All participants were fully informed about the details of the experimental procedure and gave written consent. Ethical approval for the study had been obtained from the Ethics Board of the School of Psychology at the University of Birmingham.

Materials and design

Forty-eight black-and-white line drawings of common objects were selected from a picture gallery available at the University of Birmingham (see Appendix). The database includes the Snodgrass and Vanderwart (1980) line drawings and others drawn in a similar style. Half of the objects had monosyllabic names and were on average 3.1 phonemes in length. The remaining objects had disyllabic names and were on average 5.1 phonemes in length. The disyllabic names were mono-morphemic and stressed on the first syllable. The monosyllabic and disyllabic object names were matched for frequency (mean CELEX lexical database, 2001, word

form frequencies per million words: 12.1 for monosyllabic words and 9.9 for disyllabic words).

We predicted that the durations of the gazes to the objects should vary in line with the length of the object names because it takes longer to construct the phonological form of long words than of short words. It was therefore important to ensure that the predicted Word Length effect could not be attributed to differences between the two sets in early visual–semantic processing. Therefore, we pre-tested the items in a word-picture matching task (see Jescheniak and Levelt, 1994; Stadthagen-Gonzalez et al., 2009).

The pretest was carried out with 22 undergraduate participants. On each trial, they saw one of the experimental pictures, preceded by its name or an unrelated concrete noun, which was matched to the object name for word frequency and length. Participants indicated by pressing one of two buttons whether or not the word was the name of the object. All objects occurred in the match and mismatch condition. Each participant saw half of the objects in each of the two conditions, and the assignment of objects to conditions was counterbalanced across participants. The error rate was low (2.38%) and did not differ significantly across conditions. Correct latencies between 100 and 1000 ms were analyzed in analyses of variance (ANOVAs) using length (monosyllabic vs. disyllabic) and word–picture match (match vs. mismatch) as fixed effects and either participants or items as random effects (F_1 and F_2, respectively). There was a significant main effect of word–picture match, favoring the match condition [478 ms (SE = 11 ms, by participants) vs. 503 ms (SE = 9 ms); $F_1(1, 21) = 15.5$, $p = 0.001$; $F_2(1, 46) = 4.6$, $p = 0.037$]. There was also a main effect of length, favoring the *longer* names [474 ms (SE = 11 ms) vs. 507 ms (SE = 10 ms), $F_1(1,21) = 31.1$, $p < 0.001$; $F_2(1,46) = 7.5$, $p = 0.009$]. The interaction of the two variables was not significant (both $Fs < 1$). Note that the difference in picture matching speed between the monosyllabic and disyllabic object sets was in the opposite direction than would be predicted on the basis of word length. If we observe the predicted effects of Word Length in the main experiment, they cannot be attributed to differences between the monosyllabic and disyllabic sets in early visual–conceptual processes.

The 24 objects with monosyllabic names and the 24 objects with disyllabic names were each combined into 4 sequences of 6 objects. The names in each sequence had different onset consonants, and each sequence included only one complex consonant onset. Care was taken to avoid close repetition of consonants across other word positions. The objects in each sequence belonged to different semantic categories. The pictures were sized to fit into rectangular areas of 3° × 3° visual angle and arranged in an oval with a width of 20° and a height of 15.7°.

Half of the participants named the sequences of objects with monosyllabic names and the other half named the disyllabic sequences. There were two test blocks. In each block, each display was shown on 8 successive trials, creating the total of 64 trials for every participant. The first presentation of each sequence was considered a warm-up trial and was excluded from all statistical analyses.

Apparatus. The experiment was controlled by the experimental software package NESU provided by the Max Planck Institute for Psycholinguistics, Nijmegen. The pictures were presented on a Samtron 95 Plus 19″ screen. Eye movements were monitored using an SMI EyeLink Hispeed 2D eye tracking system. Throughout the experiment, the *x*- and *y*-coordinates of the participant's point of gaze for the right eye were estimated every 4 ms. The positions and durations of fixations were computed online using software provided by SMI. Speech was recorded onto the hard disk of a GenuineIntel computer (511 MB, Linux installed) using a Sony ECM-MS907 microphone. Word durations were determined off-line using PRAAT software.

Procedure. Participants were tested individually in a sound-attenuated booth. Before testing commenced, they received written instructions and a booklet showing the experimental objects and their names. After studying these, they were asked to name the objects shown in another booklet where the names were not provided. Any errors were corrected by the experimenter. Then a practice block was run, in which the participants saw the objects on the screen one by one and named them. Then the headband of the eye tracking system was placed on the participant's head and the system was calibrated.

Speakers were told they would see sets of six objects in a circular arrangement, and that they should name them in clockwise order, starting with the object at the top. They were told that on the first presentation of a display, they should name the objects slowly and accurately, and on the seven following presentations of the same display they should aim to name the objects as quickly as possible.

At the beginning of each trial a fixation point was presented in the top position of the screen for 700 ms. Then a picture set was presented until the participant had articulated the sixth object name. The experimenter then pressed a button, thereby recording the speakers' utterance duration and removing the picture from the screen. The mean utterance duration was calculated over the eight repetitions of each set and displayed on the participant's monitor to encourage them to increase their speech rate. (These approximate utterance durations were only used to provide feedback to the participants but not for the statistical analyses of the data.) The experimenter provided additional feedback, informing the participants that their speech rate was good but encouraging them to speak faster on the next set of trials. The same procedure was used in the second block, except that the experimenter provided no further feedback. The inter-trial interval was 1250 ms.

RESULTS

Results from both experiments were analyzed with ANOVAs using subjects as a random factor, followed by linear mixed effects models and mixed logit models (Baayen et al., 2008; Jaeger, 2008). In the latter, all variables were centered before model estimates were computed. All models included participants and items (i.e., the four sequences of objects with monosyllabic names or the four sequences of objects with disyllabic names) as random effects. In Experiment 1, the fixed effects were Word Length (monosyllabic vs. disyllabic words), Block (First vs. Second Block), and Repetition. Repetition was included as a numerical predictor. Variables that did not reliably contribute to model fit were dropped. In models with interactions, only the highest-level interactions are reported below.

Error rates

Errors occurred in 7.5% of the sequences, corresponding to a rate of 1.25% of the words. Of the 115 errors, the majority were hesitations (28 errors) or anticipations of words or sounds (39 errors). The remaining errors were 9 perseverations, 6 exchanges, and 33 non-contextual errors, where object names were produced that did not appear in the experimental materials.

Inspection of the error rates showed no consistent increase or decrease across the repetitions of the picture sets. The ANOVA of the error rates yielded a significant main effect of Block [$F(1, 22) = 5.89$, $p = 0.024$] and a significant interaction of Block and Word Length [$F(1, 22) = 4.89$, $p = 0.036$]. This interaction arose because in the first block the error rate was higher for monosyllabic than for disyllabic items [11.90% (SE = 2.2%) vs. 7.74% (SE = 2.30%)], whereas the reverse was the case in the second block [4.46% (SE = 1.40%) vs. 7.74% (SE = 2.08%)]. The interaction of Block, Repetition, and Word Length was also significant [$F(6, 132) = 2.23$, $p = 0.044$]. No other effects approached significance. The mixed logit analysis of errors also showed an interaction between Block and Word Length ($\beta = 1.05$, SE = 0.44, $z = 2.41$) as well as an interaction between Word Length and Repetition ($\beta = 0.19$, SE = 0.11, $z = 1.82$). All trials on which errors occurred were eliminated from the following analyses.

Speech onset latencies

One would expect that the instruction to talk fast might affect not only speech rate, but also speech onset latencies. The average latencies for correct trials are displayed in **Figure 2**. Any latencies below 150 ms or above 1800 ms (1.1% of the data) had been excluded. In the ANOVA the main effect of Block was significant [$F(1, 22) = 87.3$, $p < 0.001$], as was the main effect of Repetition [$F(6, 132) = 13.1$, $p < 0.001$; $F(1, 22) = 34.93$, $p < 0.001$ for the linear trend]. **Figure 2** suggests longer latencies for monosyllabic than for disyllabic items, but this difference was not significant [$F(1, 22) = 1.00$, $p = 0.33$].

The best-fitting mixed effects model included main effects of Block and Repetition and an interaction between Block and Repetition ($\beta = 9$, SE = 3.41, $t = 2.67$) reflecting the fact that the effect of Repetition was stronger in the first than in the second block. There was also an interaction between Word Length and Repetition ($\beta = 9$, SE = 3.41, $t = 2.58$), as speech onsets declined over time more quickly for monosyllabic than disyllabic words. Model fit was also improved by including by-participant random slopes for Block.

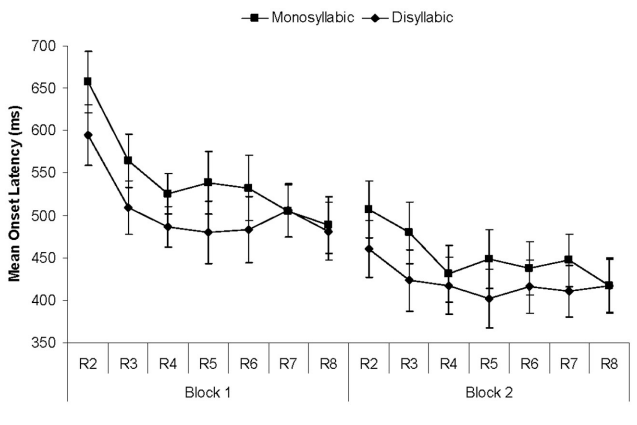


FIGURE 2 | Results of Experiment 1. Mean utterance onset latencies (ms) per block and repetition (R2 through R8) for monosyllabic and disyllabic items. Error bars show SEs.

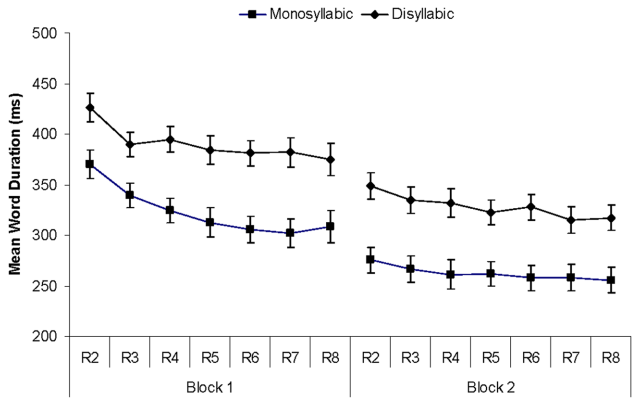


FIGURE 3 | Results of Experiment 1. Mean word durations (ms) per block and repetition (R2 through R8) for monosyllabic and disyllabic items.

Word durations

To determine how fast participants produced their utterances, we computed the average word duration for each sequence by dividing the time interval between speech onset and the offset of the last word by six[1]. As **Figure 3** shows, word durations were consistently shorter for monosyllabic than for disyllabic items; they were shorter in the second than in the first block, and they decreased across the repetitions of the sequences.

In the ANOVA, we found significant main effects of Word Length [$F(1, 22) = 15.6$, $p = 0.001$], Block [$F(1, 22) = 143.96$, $p < 0.001$], and Repetition [$F(6, 132) = 38.02$, $p < 0.001$; $F(1, 22) = 125.44$, $p < 0.001$ for the linear trend]. The interaction of Block and Repetition was also significant [$F(6, 132) = 7.22$, $p < 0.001$], as was the interaction of Word Length, Block, and Repetition [$F(6, 132) = 2.86$, $p = 0.012$]. The interaction is due to the steeper decrease in word durations in Block 1 for monosyllabic than disyllabic words. The mixed effects model showed an analogous three-way interaction ($\beta = -6$, SE = 2.48, $t = -2.29$), along with main effects of all three variables. Model fit was also improved by including by-participant random slopes for Block.

Gaze paths

To analyze the speakers' eye movements, we first determined the gaze path for each trial, i.e., established whether all objects were inspected, and in which order they were inspected. On 78.9% of

[1]The word durations included any pauses between words. Additional analyses were carried out for word durations measured from word onset to word offset and for the distribution and durations of any pauses, but, with respect to the main question of interest, these analyses provided no additional information. Both word and pause durations decreased across the repetitions of the materials and from the first to the second block.

the trials, the speakers looked at the six objects in the order of mention (simple paths). On 13.2% of the trials they failed to look at one of the objects (skips). As there were six objects in a sequence, this means that 2.2% of the objects were named without being looked at. On 4.5% of trials speakers looked back at an object they had already inspected (regressions). The remaining 3.3% of trials featured multiple skips and/or regressions.

Statistical analyses were carried out for the two most common types of gaze paths, simple paths, and paths with skips. The analysis of the proportion of simple paths yielded no significant effects. The ANOVA of the proportions of paths with skips yielded only a significant main effect of Block [$F(1, 22) = 6.77$, $p = 0.016$], with participants being less likely to skip one of the six objects of a sequence in the first than in the second block [8.1% (SE = 2.3%) vs. 21.0% (SE = 5.1%)]. The best-fitting mixed logit model included an effect of Block ($\beta = 1.04$, SE = 0.42, $t = 2.49$) and an effect of Repetition ($\beta = 0.12$, SE = 0.05, $t = 2.47$). The model also included an interaction between Block and Word Length, but including random by-participant slopes for Block reduced the magnitude of this interaction ($\beta = -0.63$, SE = 0.83, $t = -0.76$). This suggests that between-speaker differences in word durations across the two blocks largely accounted for the increase of skips on monosyllabic objects in the second block.

Gaze durations

For each trial with a simple gaze path or a single skip we computed the average gaze duration across the first five objects of the sequence. The gazes to the sixth object were excluded as participants tend to look at the last object of a sequence until the end of the trial. Durations of less than 80 ms or more than 1200 ms were excluded from the analysis (1.1% of the trials).

As **Figure 4** shows, gaze durations decreased from the first to the second block and across the repetitions within blocks, as predicted. In the first block, they were consistently longer for disyllabic than for monosyllabic items, but toward the end of the second block the Word Length effect disappeared. The ANOVA of the gaze durations yielded main effects of Block [$F(1, 22) = 21.41$, $p < 0.001$], and Repetition [$F(6, 132) = 5.39$, $p < 0.001$; $F(1, 22) = 7.35$, $p = 0.013$ for the linear trend]. The interaction of Block and Repetition was also significant [$F(6, 132) = 3.14$, $p = 0.007$], as the effect of Repetition was larger in the first than in the second block. The main effect of Word Length was marginally significant [$F(1, 22) = 3.88$, $p = 0.062$]. Finally, the three-way interaction was also significant [$F(6, 132) = 2.21$, $p = 0.05$]. Separate ANOVAs for each block showed that in the first block the main effect of Word Length was significant [$F(1, 22) = 6.39$, $p = 0.019$], as was the effect of Repetition [$F(6, 132) = 11.49$, $p < 0.001$]. In the second block, neither of the main effects nor their interaction were significant [$F < 1$ for the main effects, $F(6, 132) = 1.67$ for the interaction]. The best-fitting mixed effects model included an interaction between all three factors ($\beta = -10$, SE = 3.47, $t = -2.96$), along with three significant main effects. Including random by-participant slopes for Block improved model fit.

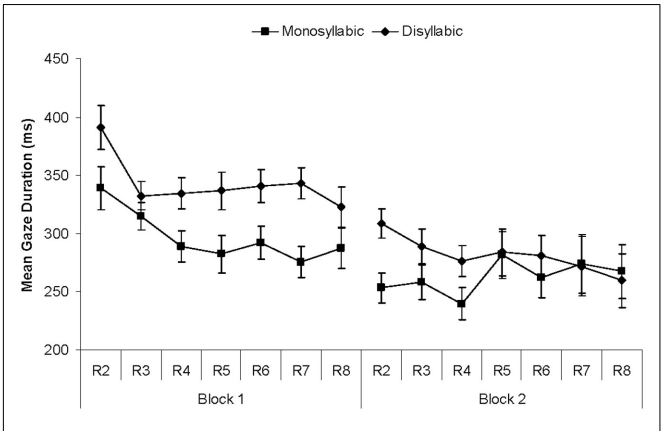


FIGURE 4 | Results of Experiment 1. Mean gaze durations (ms) per block and repetition (R2 through R8) for monosyllabic and disyllabic items.

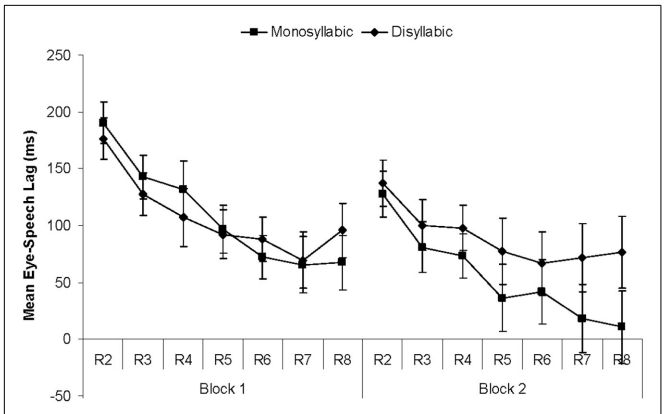


FIGURE 5 | Results of Experiment 1. Mean eye–speech lags (ms) per block and repetition (R2 through R8) for monosyllabic and disyllabic items.

Eye–speech lags

To determine the coordination of eye gaze and speech we calculated the lag between the offset of gaze to an object and the onset of its spoken name. As **Figure 5** shows, the lags decreased significantly from the first to the second block [$F(1, 22) = 11.56$, $p = 0.001$] and across the repetitions within blocks [$F(6, 132) = 21.53$, $p < 0.001$; $F(1, 22) = 66.17$, $p < 0.001$ for the linear trend]. The interaction of Block by Repetition was also significant [$F(6, 132) = 2.26$, $p < 0.05$]. Finally, the interaction of Word Length by Block approached significance [$F(1, 22) = 3.67$, $p < 0.07$]. As **Figure 5** shows, in the first block the lags for monosyllabic and disyllabic items were quite similar, but in the second block, lags were longer for disyllabic than for monosyllabic items.

The best-fitting mixed effects model included an interaction between Block and Word Length ($\beta = 35$, SE = 18, $t = 1.96$) and between Repetition and Word Length ($\beta = 7$, SE = 3, $t = 2.67$), as well as by-participant slopes for Block. Including an interaction between Block and Repetition, however, did not improve model fit [$\chi^2(1) = 1.35$, when comparing models with and without this interaction].

DISCUSSION

In Experiment 1, participants were asked to increase their speech rate across the repetitions of the materials as much as they could

without making too many errors. The analyses of participants' speech and error rates showed that they followed these instructions well: speech onset latencies and spoken word durations decreased from the first to the second block and across the repetitions within each block, while error rates remained low[2]. The speakers' eye gaze remained tightly coordinated with their speech: most of the objects were inspected, just once, shortly before they were named, and the durations of the gazes to the objects decreased along with the spoken word durations. Deviating from earlier findings, we found that the eye–speech lags decreased, rather than increased, as the speech became faster. We return to this finding in the Section "General discussion."

In addition, we observed subtle changes in the coordination of eye gaze and speech: in the second block, the objects were more likely than in the first block to be named without being fixated first, and there was a Word Length effect on gaze durations in the first but not in the second block. This indicates that in the first block the participants typically looked at each object until they had retrieved the phonological form of its name, as participants in earlier studies had done (e.g., Korvorst et al., 2006; Meyer et al., 2007), but did not do this consistently in the second block. As **Figures 4** and **5** show, in the second half of the second block, the durations of the gazes to monosyllabic and disyllabic items were almost identical, but the eye–speech lag was much longer for disyllabic than monosyllabic items. Apparently, participants disengaged their gaze from monosyllabic and disyllabic items at about the same time, perhaps as soon as the object had been recognized, but then needed more time after the shift of gaze to plan the disyllabic words and initiate production of these names.

The goal of this experiment was to explore how speakers would coordinate their eye gaze and speech when they tried to speak as fast as possible. In order to facilitate the use of a high speech rate, we presented the same pictures on several successive trials. This meant that the effects of increasing speech rate and practice were confounded. Either of those effects might be responsible for the change of the eye–speech coordination from the first to the second block. To separate the effects of practice and speech rate, a second experiment was conducted, where participants were first trained to produce the object names either at a fast or more moderate pace, and then named each display on 20 successive trials at that pace.

EXPERIMENT 2

METHOD

Participants

The experiments was carried out with 20 undergraduate students of the University of Birmingham. They had normal or corrected-to-normal vision and received either payment or course credits for participation. The participants received detailed information about the experimental procedure and gave written consent to participate.

Materials and design

The same experimental sequences were used as in Experiment 1. In addition, four training sequences were constructed, two consisting of objects with monosyllabic names and two consisting of objects with disyllabic names. All participants named the monosyllabic and the disyllabic sequences. Ten participants each were randomly assigned to the Moderate Pace and the Fast Pace group. Five participants in each of these groups were first tested on the monosyllabic sequences and then on the disyllabic sequences. For the remaining participants the order of the sequences was reversed.

Procedure

To encourage participants to adopt different speech rates, we used different presentation times for the pictures. The presentation time for the Fast Pace condition was the average speech rate over the last four repetitions of the sets in Experiment 1. This was 2150 ms for monosyllabic sequences and 2550 ms for disyllabic sequences. This corresponded to speech rates of 3.5 words/s and 2.8 words/s for the monosyllabic and disyllabic words, respectively. The moderate speech rates were 1/3 slower that the fast rates, resulting in 2850 ms of presentation time for the monosyllabic sequences (2.7 words/s) and 3350 ms for the disyllabic ones (2.1 words/s).

After receiving instructions, speakers named the objects individually, as in Experiment 1. The headband of the eye tracking system was placed on the participant's head and the system was calibrated. Speakers were instructed that they would have to maintain a constant speech rate. A tone was played coinciding with the end of the display time, which was also the time by which all six objects should have been named.

Speakers were trained on a particular pace using the training sequences. The first training sequence was presented four times while a metronome indicated the speech rate for the upcoming block. Then the metronome was switched off and the same training sequence was named eight more times. A second training sequence was then named on 12 repetitions. If the experimenter then felt that more training was required, the training sequences were repeated. When the speaker was successful in naming the training sequences at the required speed, the first four target sequences were presented. Each sequence was shown on 20 successive trials. After a short break the training and testing procedure was repeated for the second set of sequences.

RESULTS

Error rates

The first two presentations of each set were considered warm-up trials and were excluded from all analyses. 11.28% of the remaining trials included one more errors. Of the 333 errors, most were hesitations (131 errors), anticipations (112 errors), or incorrect object names (67 errors), which were nouns that were not names of any of the objects in the current display. The remaining errors were 5 perseverations, 10 exchanges, and 8 combinations of errors. The error rates for the different experimental conditions are shown in **Figure 6**. In the ANOVA only the main effect of Repetition was significant [$F(17, 306) = 3.38$, $p < 0.001$] with errors becoming less frequent across the repetitions $F(1, 18) = 23.14$, $p < 0.001$ for the linear trend). The mixed logit analysis of errors

[2]The speech planning model proposed by Dell et al. (1997) predicts that the proportion of anticipatory errors in a set of movement errors should increase as the overall error rate decreases. In our experiment, the total number of movement errors in the first and second block was almost the same (25 vs. 23 errors), but the proportion of anticipatory errors was much higher in the second than in the first block [96% (1 out 23) vs. 68% (8 out of 25)]. This finding is in line with the assumption of the model that practice strengthens the links between the plan and upcoming units.

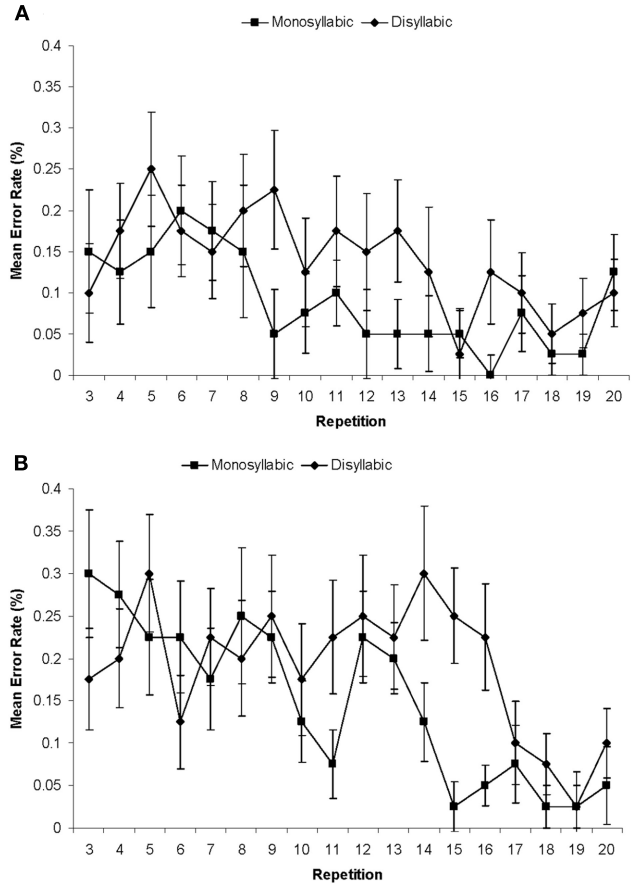


FIGURE 6 | Results of Experiment 2. Mean error rates per sequence for monosyllabic and disyllabic items on repetitions 3 through 20 of the materials in the Moderate Pace Group **(A)** and the Fast Pace Group **(B)**.

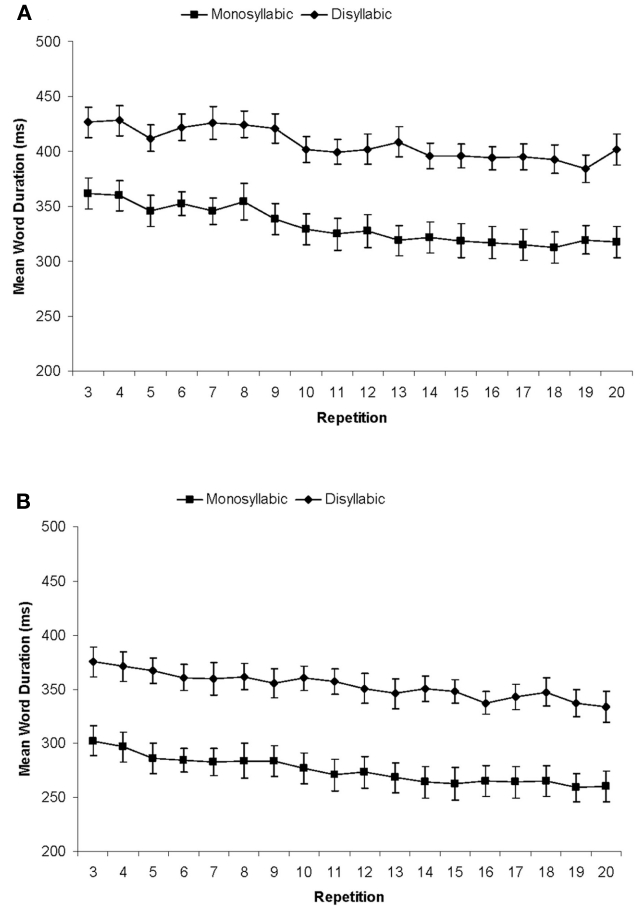


FIGURE 7 | Results of Experiment 2. Mean word durations (ms) for monosyllabic and disyllabic items on repetitions 3 through 20 in the Moderate Pace Group **(A)** and the Fast Pace Group **(B)**.

showed an interaction between Repetition and Word Length ($\beta = 0.06$, SE $= 0.02$, $z = 2.30$), as the number of errors on disyllabic sequences dropped to the level of errors on monosyllabic sequences over time. Model fit was improved by including random by-participant slopes for Word Length.

Speech onset latencies

The participants were instructed to maintain specific speech rates across the repetitions of the materials. The Fast Pace speakers timed their utterances well and completed the monosyllabic sequences on average after 2017 ms (target: 2150 ms) and the disyllabic sequences after 2506 ms (target: 2550 ms). The Moderate Pace completed the monosyllabic sequences after 2425 ms (target: 2850 ms) and the disyllabic sequences after 2908 ms (target: 3350 ms). This was faster than expected, but there was still a considerable difference in utterance completion time to the Fast Pace group.

The analyses of the speech onset latencies (excluding all errors and 1.1% of the trials with latencies below 150 ms) only yielded a main effect of Word Length, with sequences of monosyllabic words being initiated faster than sequences of disyllabic words [means: 456 ms (SE = 19 ms) vs. 490 ms (SE = 22 ms); $F(1, 18) = 7.66$; $p = 0.013$]. The best-fitting mixed effects model included a marginal effect of Word Length ($\beta = 34.12$, SE $= 18.69$, $t = 1.83$) and by-participant random slopes for this factor.

Word durations

Analyses of variance of word durations showed the expected effects of Word Length [$F(1, 18) = 641.31$; $p < 0.001$] and Pace [$F(1, 18) = 10.64$; $p < 0.001$]. The main effect of Repetition was also significant [$F(17, 306) = 22.06$, $p < 0.001$]. As **Figure 7** shows, word durations decreased across the repetitions of the materials, yielding a significant linear trend [$F(1, 18) = 72.41$, $p < 0.001$]. There were no significant interactions.

In the mixed effects linear model, all three factors contributed to model fit as additive effects (Word Length: $\beta = 83$, SE $= 9$, $t = 8.82$; Pace: $\beta = -51$, SE $= 15$, $t = -3.29$; and Repetition: $\beta = -2$, SE $= 0.28$, $t = -8.54$). Model fit was also improved by including random by-participant slopes for Word Length and Repetition, as well as random by-item slopes for Pace.

Gaze paths

As in Experiment 1, the participants usually (on 78.51% of the trials) looked at all objects in the order of mention (simple gaze paths). On 17.87% of trials they skipped one of the six objects. They produced regressions (looking back at an object they had already inspected) on only 0.2% of trials. The remaining 3.42% of trials included multiple skips and/or regressions.

In the analysis of the proportions of simple paths, only the main effect of Pace was significant [$F(1, 18) = 7.50$, $p = 0.013$] with the speakers using the Moderate Pace being more likely to fixate upon all objects than the speakers using the Fast Pace [means: 85.51% (SE = 0.62%) vs. 71.30% (SE = 0.62%)]. The best-fitting mixed effects model included a three-way interaction ($\beta = -0.09$, SE = 0.04, $z = -2.17$), with a marginally reliable main effect of Pace and interaction between Pace and Repetition. This pattern arose because speakers using the Moderate Pace, but not the speakers using the Fast Pace, were more likely to fixate objects with disyllabic names than objects with monosyllabic names at later repetitions of the picture sequences. Model fit was also improved by including random by-participant slopes for Word Length and Repetition as well as random by-item slopes for Pace and Repetition.

The analysis of the proportion of trials with skips yielded a complementary pattern: there was only a significant main effect of Pace [$F(1, 18) = 7.74$, $p = 0.012$], with speakers using the fast pace skipping one of the objects more often (on 23.77% of the trials) than speakers using the moderate pace (10.18% of the trials, SE = 0.45% for both groups).

The best-fitting mixed effects model included a main effect of Pace, a two-way interaction between Word Length and Repetition, as well as a three-way interaction ($\beta = 0.12$, SE = 0.05, $z = 2.38$). The rate at which speakers skipped pictures increased over time, but speakers using the Moderate Pace were less likely to skip pictures with disyllabic than with monosyllabic names. Model fit was also improved by including random by-participant slopes for Word Length and Repetition.

Gaze durations

Gaze durations were calculated in the same way as for Experiment 1. The statistical analysis showed a significant main effect of Pace [$F(1, 18) = 15.47$, $p = 0.001$], as gazes were shorter at the Fast than the Moderate pace (see **Figure 8**). The main effect of Repetition was also significant [$F(17, 306) = 4.19$, $p < 0.002$], with gaze durations decreasing across the repetitions of the sequences [$F(1, 18) = 11.85$, $p = 0.003$ for the linear trend]. Finally, the main effect of Word Length was significant [$F(1, 21) = 123.87$, $p < 0.001$], with gaze durations being longer, by 53 ms, in the disyllabic than in the monosyllabic sets. There were no significant interactions. The best-fitting mixed effects model included all three factors as additive effects (Word Length: $\beta = 48$, SE = 11, $t = 4.42$; Pace: $\beta = -55$, SE = 13, $t = -4.12$; and Repetition: $\beta = -1.24$, SE = 0.51, $t = -2.42$). Model fit was also improved by including random by-participant slopes for Word Length and Repetition.

Eye–speech lags

In contrast to gaze durations, eye–speech lags were not affected by the Pace, $F < 1$ (see **Figure 9**). There was a significant main effect of Word Length for the lags [$F(1, 18) = 5.44$, $p = 0.032$], which were longer by 19 ms for disyllabic than for monosyllabic sets. Lags decreased significantly across the repetitions of the materials [$F(17, 306) = 26.94$, $p < 0.001$ for the main effect; $F(1, 18) = 99.38$, $p < 0.001$ for the linear trend].

The best-fitting mixed effects model included Repetition as an additive effect as well as an interaction between Repetition and Word Length ($\beta = 2$, SE = 0.67, $t = 3.40$), as lags for monosyllabic object names decreased more quickly than for disyllabic object names. Model fit was also improved by including random by-participant slopes for Word Length and Repetition.

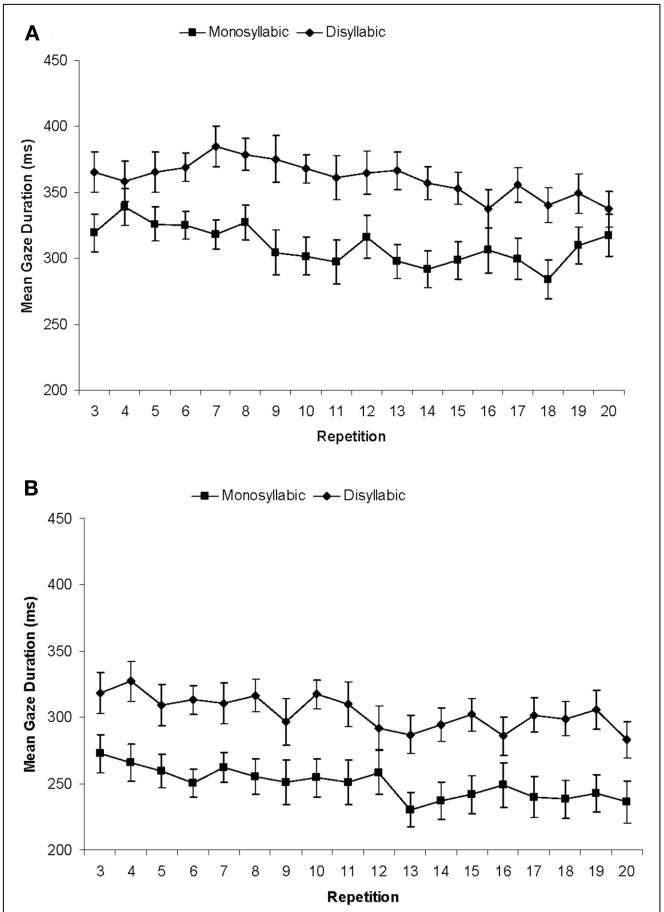


FIGURE 8 | Results of Experiment 2. Mean gaze durations (ms) for monosyllabic and disyllabic items on repetitions 3 through 20 in the Moderate Pace Group **(A)** and the Fast Pace Group **(B)**.

DISCUSSION

In Experiment 1, the effects of practice and speech rate on speech-to-gaze alignment were confounded, as participants practiced naming the object sets at increasing speed. In Experiment 2 we aimed to separate these effects by asking participants to name the object sequences repeatedly at a fixed moderate or faster pace.

The comparisons of the Fast vs. Moderate Pace group showed that the difference in speech rate was accompanied by differences in gaze durations, while there was no difference in the eye–speech

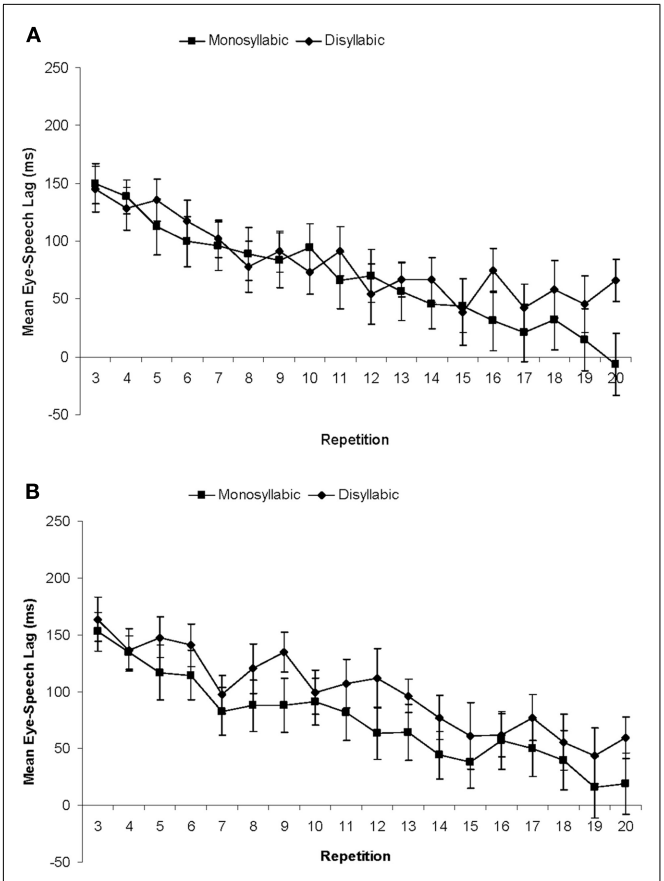


FIGURE 9 | Results of Experiment 2. Mean lags (ms) for monosyllabic and disyllabic items on repetitions 3 through 20 in the Moderate Pace Group **(A)** and the Fast Pace Group **(B)**.

lags. Thus, speakers adopting different speech rates differed in how much time they spent looking at and attending to each object, but they did not differ in their planning span.

Evaluating the effects of practice was complicated by the fact that participants, contrary to the instructions, slightly increased their speech rate across the repetitions of the sequences. This increase in speech rate was accompanied by a small decrease in gaze durations and a more substantial decrease in the lags. Apparently those processes carried out before the shift of gaze from one object to the next as well as those carried out after the shift of gaze became faster as participants became more familiar with the materials, and this resulted in the unintended increase in speech rate. The reasons why the lags decreased more than the gaze durations are not clear. One possibility is that focused attention was required to counteract practice effects arising from the repetition of the materials; as soon as attention turned from the current object to the next object, the remaining planning processes were completed at a default pace which increased as the materials became more familiar.

In Experiment 1, we had observed an effect of Word Length on gaze durations in the first but not in the second test block. By contrast, in Experiment 2 the Word Length effect on gaze durations was maintained across the entire experiment, demonstrating that participants consistently fixated upon each object until they had retrieved the phonological form of its name. Moreover, we found a significant Word Length effect for the eye–speech lag. This effect reflects the fact that the processes the speakers carried out after the shift of gaze – i.e., phonetic and articulatory encoding – took more time for the longer words. No length effect on the lags had been seen in Experiment 1. A possible account of this difference between the experiments is that in Experiment 2, participants typically planned both syllables of the disyllabic items before speech onset, whereas in Experiment 1 they often planned only one syllable before the onset of articulation (Meyer et al., 2003; Damian et al., 2010). In line with this account, Experiment 2 also yielded an effect of Word Length on utterance onset latencies, which had been absent in Experiment 1. This might reflect that in Experiment 2 the participants usually planned the full phonological and phonetic form of the first object name before beginning to speak, whereas in Experiment 1 they often initiated the utterances earlier (for further discussions of word length effects on utterance onset latencies see Bachoud-Lévi et al., 1998; Griffin, 2003, and Meyer et al., 2003). Thus, the different instructions affected not only the speech rates, but also led the participants of Experiment 2 to adopt a more careful, deliberate speech style.

GENERAL DISCUSSION

The goal of these studies was to explore how speakers' gaze-to-speech alignment would change when they used different speech rates. We found that, at all speech rates, the speakers' eye movements were systematically coordinated with their overt speech: they fixated upon most of the objects before naming them, and shorter spoken word durations were accompanied by shorter gazes to the objects. As explained in the Introduction, there is strong evidence that gaze durations reflect on the times speakers take to recognize the objects and to plan their names. Therefore the decreasing gaze durations observed with increasing speech rates show that speakers spend less time attending to each of the objects and preparing their names when they speak fast than when they speak more slowly.

In our study, the eye–speech lag, the time between the shift of gaze from one object to the next and the onset of the name of the first object, was not systematically affected by speech rate. There was no evidence that speakers would plan their utterances further, or less far, ahead when they used different speech rates. This result is consistent with findings reported by Spieler and Griffin (2006) and Belke and Meyer, 2007, Experiment 1). It contrasts with results obtained by Griffin (2003), Mortensen et al. (2008), and Belke and Meyer (2007), who found that slower speech rates were associated with shorter eye–speech lags than faster speech rates. The reasons for this difference in the results are not clear. Griffin (2003) and Mortensen et al. (2008) compared the eye–speech coordination of speakers differing in their habitual speech rate. It is not known why the speakers differed in speech rate. One possibility is that the slower speakers initiated the shift of gaze to a new object later, relative to the speech planning processes, than the faster speakers: whereas the faster speakers directed their eye gaze and attention to a new object as soon as they had retrieved the phonological form of the name of the current object, the slower speakers only initiated the shift a little later, after completing part of the phonetic

or articulatory encoding as well. This would yield longer gazes and shorter lags in slower compared to faster speakers. The slower speakers might have used such a more cautious processing strategy because they monitored their speech more carefully or tried to minimize the interference from new visual input (Griffin, 2003, 2004; Belke and Meyer, 2007). Similarly, when the speakers in Belke and Meyer's study were asked to use a very slow, rather than moderate speech rate, they may have altered the criterion for initiating the shifts of gaze from one object to the next because maintaining a very slow speech rate might require close attention to the phonetic and articulatory planning processes.

In the present study we did not observe systematic changes of the eye–speech lags with different speech rates, but we did see that the gaze-to-speech coordination was much tighter in Experiment 2 than in Experiment 1. This is evidenced clearly by the effects of name length on gaze durations and eye–speech lags found only in Experiment 2. The speakers of Experiment 2 were explicitly asked to pay attention to an aspect of the form of their utterances, the speech rate. Maintaining the prescribed speech rate was difficult, as shown by the fact that the Moderate Pace group consistently produced their utterances too fast, and that both groups increased their speech rate across the repetitions of the materials. As the names of the objects became available more and more quickly with increasing practice, it would probably have been easier for participants to increase their speech rate than to counteract the practice effects and maintain a constant speech rate. The systematic alignment of the shifts of eye gaze and attention with the completion of phonological planning may have been a way of supporting and monitoring the regular timing of the utterances. By contrast, the participants of Experiment 1 could produce the object names as soon as they had been planned, and no monitoring of speech rate was required. Since the production of the sequences of object names became less demanding with practice, attending to the objects until their names had been planned became less beneficial, and therefore speakers often moved their eyes to the next object before they had retrieved the phonological form of the present object.

CONCLUSION

What do speakers do when they use different speech rates? Our study showed that eye gaze and speech output remained well coordinated across a range of speech rates: regardless of the speech rate, speakers look at most of the objects they name, and when they spoke faster the durations of the gazes to the objects decreased. This indicates that speakers spent less time planning each object name. When they were required to maintain a fixed speech rate, their eye gaze-to-speech coordination was very tight, with the shift of gaze being initiated after the name of each object had been phonologically encoded. This might be because maintaining a fixed speech rate is difficult, especially for well-practiced utterances, and requires careful monitoring of the speech output. When speakers aimed to speak as fast as they could, they initially still looked at, and attended to, each object until they had retrieved the phonological form of its name, but later moved their eyes earlier, perhaps as soon as they had identified the object. Together, the results suggest that looking at each object until its name has been planned to the level of the phonological form is a well established default strategy. Speakers use it because attending to the objects facilitates the recognition of the objects and the retrieval of their names. However, speakers can deviate from this strategy when they aim to monitor their speech very carefully, or when the utterances they produce are highly familiar and close monitoring is not required.

ACKNOWLEDGMENTS

This research was partly funded by ESRC grant R000239659 to Antje S. Meyer and Linda R. Wheeldon. It was carried out under the auspices of the University of Birmingham, UK.

REFERENCES

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412.

Bachoud-Lévi, A. C., Dupoux, E., Cohen, L., and Mehler, J. (1998). Where is the length effect? A cross-linguistic study of speech production. *J. Mem. Lang.* 39, 331–346.

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.

Belke, E., and Meyer, A. S. (2007). Single and multiple object naming in healthy aging. *Lang. Cogn. Process.* 22, 1178–1210.

Bock, K., Irwin, D. E., Davidson, D. J., and Levelt, W. J. M. (2003). Minding the clock. *J. Mem. Lang.* 48, 653–685.

CELEX lexical database. (2001). Available at: http://celex.mpi.nl

Cook, A. E., and Meyer, A. S. (2008). Capacity demands of word production: new evidence from dual-task experiments. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 886–899.

Damian, M. F., Bowers, J. S., Stadthagen-Gonzalez, H., and Spalek, K. (2010). Does word length affect speech onset latencies in single word production? *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 892–905.

Dell, G. S., Burger, L. K., and Svec, W. R. (1997). Language production and serial order: a functional analysis and a model. *Psychol. Rev.* 104, 123–147.

Deubel, H., and Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Res.* 36, 1827–1837.

Eimer, M., van Velzen, J., Gherri, E., and Press, C. (2007). ERP correlates of shared control mechanisms involved in saccade preparation and in covert attention. *Brain Res.* 1135, 145–166.

Ferreira, V. S., and Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 1187–1199.

Gleitman, L. R., January, D., Nappa, R., and Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *J. Mem. Lang.* 57, 544–569.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition.* 82, B1–B14.

Griffin, Z. M. (2003). A reversed length effect in coordinating the preparation and articulation of words in speaking. *Psychon. Bull. Rev.* 10, 603–609.

Griffin, Z. M. (2004). "Why look? Reasons for eye movements related to language production," in *The Interface of Language, Vision, and Action: What We Can Learn from Free-Viewing Eye Tracking*, eds J. M. Henderson and F. Ferreira (New York: Psychology Press), 213–247.

Griffin, Z. M., and Bock, K. (2000). What the eyes say about speaking. *Psychol. Sci.* 11, 274–279.

Griffin, Z. M., and Oppenheimer, D. M. (2006). Speakers gaze at objects while preparing intentionally inaccurate names for them. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 943–948.

Hartsuiker, R. J. (2006). Are speech error patterns affected by a monitoring bias? *Lang. Cogn. Process.* 21, 856–891.

Hartsuiker, R. J., Pickering, M. J., and De Jong, N. (2005). Semantic and phonological context effects in speech error repair. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 921–932.

Irwin, D. E. (2004). "Fixation location and fixation duration as indices of cognitive processes," in *The Interface of Language, Vision, and Action: What We Can Learn from Free-Viewing Eye Tracking*, eds J. M. Henderson and F. Ferreira (New York: Psychology Press), 105–133.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards mixed logit models. *J. Mem. Lang.* 59, 434–446.

Jescheniak, J. D., and Levelt, W. J. M. (1994). Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 824–843.

Korvorst, M., Roelofs, A., and Levelt, W. J. M. (2006). Incrementality in naming and reading complex numerals: evidence from eyetracking. *Q. J. Exp. Psychol.* 59, 296–311.

Meyer, A. S., Belke, E., Häcker, C., and Mortensen, L. (2007). Regular and reversed word length effects in picture naming. *J. Mem. Lang.* 57, 210–231.

Meyer, A. S., Roelofs, A., and Levelt, W. J. M. (2003). Word length effects in picture naming: the role of a response criterion. *J. Mem. Lang.* 47, 131–147.

Meyer, A. S., Sleiderink, A. M., and Levelt, W. J. M. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition* 66, B25–B33.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wegner, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks. A latent variable analyses. *Cogn. Psychol.* 41, 49–100.

Mortensen, L., Meyer, A. S., and Humphreys, G. W. (2008). Speech planning during multiple-object naming: effects of aging. *Q. J. Exp. Psychol.* 61, 1217–1238.

Oomen, C. C. E., and Postma, A. (2002). Limitations in processing resources and speech monitoring. *Lang. Cogn. Process.* 17, 163–184.

Posner, M. I., and Petersen, S. E. (1990). The attention system of the human brain. *Annu. Rev. Neurosci.* 13, 25–42.

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition* 77, 97–131.

Roelofs, A. (2007). Attention and gaze control in picture naming, word reading, and word categorization. *J. Mem. Lang.* 57, 232–251.

Roelofs, A. (2008a). Attention, gaze shifting, and dual-task interference from phonological encoding in spoken word planning. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1580–1598.

Roelofs, A. (2008b). Tracing attention and the activation flow in spoken word planning using eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 353–367.

Slevc, L. R., and Ferreira, V. S. (2006). Halting in single-word production: a test of the perceptual-loop theory of speech monitoring. *J. Mem. Lang.* 54, 515–540.

Snodgrass, J. G., and Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol. Hum. Learn. Mem.* 6, 174–215.

Spieler, D. H., and Griffin, Z. M. (2006). The influence of age on the time course of word preparation in multiword utterances. *Lang. Cogn. Process.* 21, 291–321.

Stadthagen-Gonzalez, H., Damian, M. F., Pérez, M. A., Bowers, J. S., and Marín, J. (2009). Name-picture verification as a control measure for object naming: a task analysis and norms for a large set of pictures. *Q. J. Exp. Psychol.* 62, 1581–1597.

Wühr, P., and Frings, C. (2008). A case for inhibition. Visual attention suppresses the processing of irrelevant objects. *J. Exp. Psychol. Gen.* 137, 116–130.

Wühr, P., and Waszak, F. (2003). Object-based attentional selection can modulate the Stroop effect. *Mem. Cognit.* 31, 983–994.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 July 2011; accepted: 02 February 2012; published online: 20 February 2012.

*Citation: Meyer AS, Wheeldon L, van der Meulen F and Konopka A (2012) Effects of speech rate and practice on the allocation of visual attention in multiple object naming. Front. Psychology **3**:39. doi: 10.3389/fpsyg.2012.00039*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

APPENDIX

MONOSYLLABIC SETS

Lamp coin rope bat straw pie; pin toe spoon leaf bow rat; owl mask web corn sword brush; kite doll tap sock whale globe.

DISYLLABIC SETS

Lemon toilet spider pencil coffin basket; saddle bucket penguin ladder whistle carrot; barrel wardrobe monkey statue rabbit garlic; sausage dragon robot tortoise candle orange.