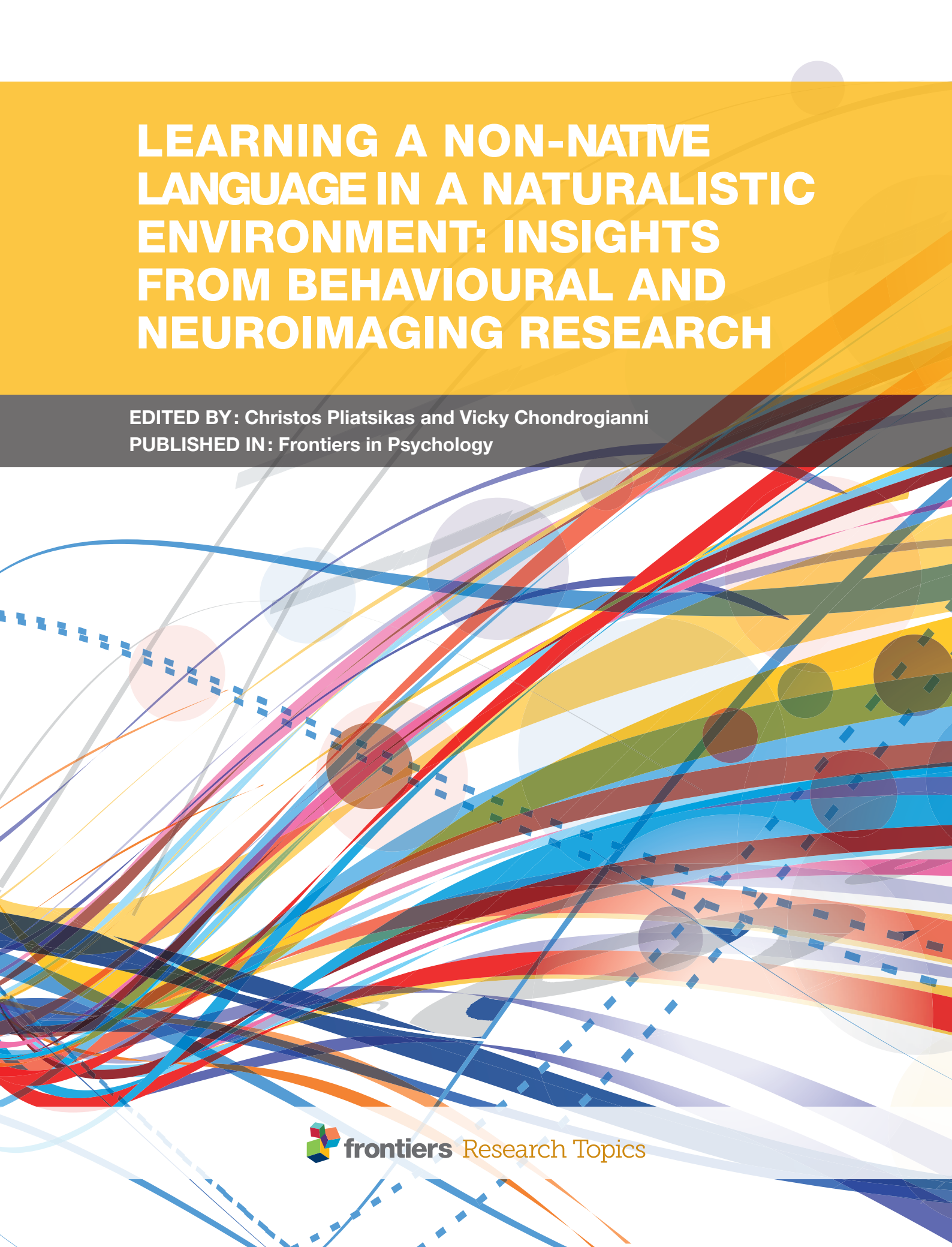
LEARNING A NON-NATIVE LANGUAGE IN A NATURALISTIC ENVIRONMENT: INSIGHTS FROM BEHAVIOURAL AND NEUROIMAGING RESEARCH

EDITED BY: Christos Pliatsikas and Vicky Chondrogianni
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-639-5
DOI 10.3389/978-2-88919-639-5

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**

LEARNING A NON-NATIVE LANGUAGE IN A NATURALISTIC ENVIRONMENT: INSIGHTS FROM BEHAVIOURAL AND NEUROIMAGING RESEARCH

Topic Editors:
Christos Pliatsikas, University of Kent, UK
Vicky Chondrogianni, University of Edinburgh, UK

It is largely accepted in the relevant literature that successful learning of one or more non-native languages is affected by a number of factors that are independent of the target language(s) per se; these factors include the age of acquisition (AoA) of the target language(s), the type and amount of formal instruction the learners have received, as well as the amount of language use that the learners demonstrate. Recent experimental evidence suggests that one crucial factor for efficient native-like performance in the non-native language is the amount of naturalistic exposure, or immersion, that the learners receive to that language. This can be broadly defined as the degree to which language learners use their non-native language outside the classroom and for their day-to-day activities, and usually presupposes that the learners live in an environment where their non-native language is exclusively or mostly used.

Existing literature has suggested that linguistic immersion can be beneficial for lexical and semantic acquisition in a non-native language, as well as for non-native morphological and syntactic processing. More recent evidence has also suggested that naturalistic learning of a non-native language can also have an impact on the patterns of brain activity underlying language processing, as well as on the structure of brain regions that are involved, expressed as changes in the grey matter structure.

This Research Topic brings together studies on the effects of learning and speaking a non-native language in a naturalistic environment. These include more efficient or "native-like" processing in behavioural tasks tapping on language (lexicon, morphology, syntax), as well as changes in the brain structure and function, as revealed by neuroimaging studies.

Citation: Pliatsikas, C., Chondrogianni, V., eds. (2015). Learning a Non-Native Language in a Naturalistic Environment: Insights from Behavioural and Neuroimaging Research. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-639-5

Table of Contents

frontiers in Psychology

EDITORIAL
published: 17 July 2015
doi: 10.3389/fpsyg.2015.01009

Editorial: Learning a non-native language in a naturalistic environment: insights from behavioral and neuroimaging research

Christos Pliatsikas[1]* *and Vicky Chondrogianni*[2]

[1] School of Psychology, University of Kent, Canterbury, UK, [2] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

Keywords: bilingualism, second language acquisition, immersion

Research on bilingualism has boomed in the past two decades. The processes by which a second language is acquired and processed has been investigated via linguistic, psycholinguistic, and neurolinguistic perspectives, focusing not only on second language (L2) acquisition and processing, but also the effects it might have on cognition and brain structure and function (Bialystok et al., 2012). More recent studies have focused on the effects of experience-based factors on L2 acquisition and processing (Dussias and Piñar, 2009); for example, several studies have increasingly focused on how L2 processing is affected by the active and continuous use of L2, or immersion, whether it becomes native-like, and which language domains are particularly affected (Dussias and Sagarra, 2007; Pliatsikas and Marinis, 2013). The present E-Book is a collection of recent studies that demonstrate the effects of immersive L2 learning in lexical, phonological and morpho-syntactic processing, while at the same time discusses the potential effects of immersive non-native acquisition on the structure of the bilingual brain.

Several studies in this E-Book have focused on morpho-syntactic processing by immersed late L2 learners. In an ERP study, Carrasco-Ortiz and Frenck-Mestre (2014) showed that highly proficient L2 learners of French with limited immersion (5–6 months) were native-like in their sensitivity of detecting verbal inflectional errors. This sensitivity was enhanced in the presence of phonological cues to the errors, but was also dependent on the L2 learner's overall proficiency. Further evidence in the domain of morpho-syntax was provided in an ERP experiment by Meulman et al. (2014), who demonstrated that immersed (5 years) late Romance learners of Dutch were native-like in detecting auditorily-presented verb agreement violations in non-finite verbs, but not gender violations. This demonstrated that there might be limits to how native-like L2 processing can be, but these limits are specific to the grammatical construction under investigation.

In two behavioral masked lexical priming experiments and in an ERP study with advanced Spanish and German late L2 learners of English, De Cat et al. (2015) showed that lexically transparent noun-noun compounds (NNCs) such as *moon dust* are processed combinatorially by advanced non-native speakers similarly to native speakers; however, sensitivity to word order violations within the NNCs was modulated by the learners' L1.

In an acceptability judgment task, Parafita Couto et al. (2015) examined the interaction between word order and focus in the context of unaccusative (e.g., *arrive*) and unergative (e.g., *walk*) verbs in Spanish in a group of English late L2 learners of Spanish with extensive naturalistic exposure to L2 input. Immersed late L2 learners accepted different word order patterns depending on the focus context; however, they failed to distinguish between unaccusative and unergative verbs, and the ability to do so was a function of the verb's frequency rather than its categorical classification on

OPEN ACCESS

Edited and reviewed by:
Manuel Carreiras, Basque Center on Cognition, Brain and Language, Spain

***Correspondence:**
Christos Pliatsikas, c.pliatsikas@reading.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *29 June 2015*
Accepted: *06 July 2015*
Published: *17 July 2015*

Citation:
Pliatsikas C and Chondrogianni V (2015) Editorial: Learning a non-native language in a naturalistic environment: insights from behavioral and neuroimaging research. Front. Psychol. 6:1009. doi: 10.3389/fpsyg.2015.01009

the basis of unaccusativity. At the same time, L2 learners were less categorical in their judgments compared to monolingual speakers.

In terms of lexical recognition, in two behavioral experiments, Casaponsa et al. (2014) demonstrated that immersed balanced and unbalanced Spanish-Basque bilinguals were equally efficient in recognizing L2-specific bigrams, suggesting that bilingual immersion can lead to native-like orthographic processing; however, these effects were modulated by the participants' L2 proficiency.

Zinszer et al. (2014) tested Chinese-English bilinguals in China and in the US on a lexical categorization task and examined which L2 learner's language history variables (length of immersion, L2 training, age of L2 onset, and code-switching patterns) and language variables (e.g., native speaker agreement on picture naming) predict performance on this task. The authors reported that words with high name agreement and few alternate names elicited high performance; at the same time, immersion, age of L2 onset and code-switching patterns contributed positively to learners' performance, whereas years of L2 training had a negative impact on task performance.

The effects of exposure to naturalistic L2 input on vocabulary learning were examined in two studies by Dahl and Vulchanova (2014) and by Vulchanova et al. (2015). Dahl and Vulchanova examined whether providing naturalistic L2 exposure within a standard school curriculum influences comprehension of vocabulary in two groups of 6-year-old Norwegian-speaking children. After 8 months of exposure, the group that received naturalistic input to English outside the classroom setting but within the school context outperformed on vocabulary learning the group that was only exposed to English within the classroom setting. This suggests that increased exposure to the L2 can lead to a significant increase in receptive vocabulary at this young age even after a short period.

Vulchanova et al. (2015) examined short- and long-term memory effects of first language (L1) and L2 subtitles on text comprehension and vocabulary learning in two groups of adolescent Norwegian learners of English. Short-term effects of L1 and L2 subtitles on text comprehension were found in both groups. These effects were modulated by vocabulary knowledge in the younger group of L2 learners and by knowledge of grammar in the older L2 group. There were no long-term effects in either group on vocabulary learning as measured through a word definition task and lexical decision task. Participants' extracurricular activities such as reading and writing in the L2, exposure to L2 media and games also emerged as significant predictors of the L2 learners' comprehension abilities.

In terms of phonological processing, Gor (2014) demonstrated that heritage English-Russian speakers (early naturalistic interrupted learners) of high proficiency in Russian, were equally efficient to native speakers of Russian in processing speech in noise. This demonstrated the early benefits of immersed L2 learning, which appear to persevere even when immersion is interrupted.

Although the existing behavioral and ERP literature appears to argue for substantial effects of immersion on bilinguals' performance, its effects on brain structure are proven more difficult to capture and describe. In an opinion article, Sharwood Smith (2014) discusses the issues in combining linguistic, psychological and neuroimaging approaches in the search for a unified theory of bilingual processing. In reviewing the neurolinguistic literature, Stein et al. (2014) argue that the reported structural effects of bilingualism on the gray matter (GM) and white matter (WM) of the brain cannot be safely attributed to the type or amount of L2 immersion, although it appears that immersion is more likely to have an impact on the WM (see also Pliatsikas et al., 2015). The effects of bi-/multilingualism on the GM are further demonstrated in a structural MRI study by Kaiser et al. (2015). In this study, possibly the first of its kind on multilinguals, it is suggested that successive L2 learning leads to more extended changes in GM compared to early simultaneous language learning. This effect persists even in individuals that learn a third language later in life, suggesting that early immersive bilingualism might lead to more effective synaptic connectivity for language learning, which in turn leads to less profound structural changes during late learning of additional languages.

Taken together, the papers in this E-book demonstrate the role and the importance of experienced-based factors, and especially linguistic immersion, for the acquisition and processing of a second or a third language. We hope that this E-book will inspire researchers to pay particular attention to the environmental factors that shape the linguistic experiences of their non-native participants, and to present comprehensive descriptions of their groups' linguistic background, including detailed information about their bi-/multilingual immersion.

References

Bialystok, E., Craik, F. I. M., and Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends Cogn. Sci.* 16, 240–250. doi: 10.1016/j.tics.2012.03.001

Carrasco-Ortiz, H., and Frenck-Mestre, C. (2014). Phonological and orthographic cues enhance the processing of inflectional morphology. ERP evidence from L1 and L2 French. *Front. Psychol.* 5:888. doi: 10.3389/fpsyg.2014.00888

Casaponsa, A., Carreiras, M., and Duñabeitia, J. A. (2014). Discriminating languages in bilingual contexts: the impact of orthographic markedness. *Front. Psychol.* 5:424. doi: 10.3389/fpsyg.2014.00424

Dahl, A., and Vulchanova, M. D. (2014). Naturalistic acquisition in an early language classroom. *Front. Psychol.* 5:329. doi: 10.3389/fpsyg.2014.00329

De Cat, C., Klepousniotou, E., and Baayen, R. H. (2015). Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English. *Front. Psychol.* 6:77. doi: 10.3389/fpsyg.2015.00077

Dussias, P. E., and Piñar, P. (2009). "Sentence Parsing in L2 Learners: Linguistic and Experience-based Factors," in *The New Handbook of Second Language Acquisition*, eds W. C. Ritchie and T. K. Bathia (Bingley: Emerald Group Publishing), 295–318.

Dussias, P. E., and Sagarra, N. (2007). The effect of exposure on syntactic parsing in Spanish–English bilinguals. *Biling. Lang. Cogn.* 10, 101–116. doi: 10.1017/S1366728906002847

Gor, K. (2014). Raspberry, not a car: context predictability and a phonological advantage in early and late learners' processing of

speech in noise. *Front. Psychol.* 5:449. doi: 10.3389/fpsyg.2014.01449

Kaiser, A., Eppenberger, L. S., Smieskova, R., Borgwardt, S., Kuenzli, E., Radue, E.-W., et al. (2015). Age of second language acquisition in multilinguals has an impact on gray matter volume in language-associated brain areas. *Front. Psychol.* 6:638. doi: 10.3389/fpsyg.2015.00638

Meulman, N., Stowe, L. A., Sprenger, S. A., Bresser, M., and Schmid, M. S. (2014). An ERP study on L2 syntax processing: when do learners fail? *Front. Psychol.* 5:1072. doi: 10.3389/fpsyg.2014.01072

Parafita Couto, M. C., Mueller Gathercole, V. C., and Stadthagen-Gonzalez, H. (2015). Interface strategies in monolingual and end-state L2 Spanish grammars are not that different. *Front. Psychol.* 5:1525. doi: 10.3389/fpsyg.2014.01525

Pliatsikas, C., and Marinis, T. (2013). Processing empty categories in a second language: when naturalistic exposure fills the (intermediate) gap. *Biling. Lang. Cogn.* 16, 167–182. doi: 10.1017/S136672891200017X

Pliatsikas, C., Moschopoulou, E., and Saddy, D. (2015). The effects of bilingualism on grey and white matter structure. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1334–1337. doi: 10.1073/pnas.1414183112

Sharwood Smith, M. A. (2014). In search of conceptual frameworks for relating brain activity to language function. *Front. Psychol.* 5:716. doi: 10.3389/fpsyg.2014.00716

Stein, M., Winkler, C., Kaiser, A., and Dierks, T. (2014). Structural brain changes related to bilingualism: does immersion make a difference? *Front. Psychol.* 5:1116. doi: 10.3389/fpsyg.2014.01116

Vulchanova, M., Aurstad, L. M. G., Kvitnes, I. E. N., and Eshuis, H. (2015). As naturalistic as it gets: subtitles in the English classroom in Norway. *Front. Psychol.* 5:1510. doi: 10.3389/fpsyg.2014.01510

Zinszer, B. D., Malt, B. C., Ameel, E., and Li, P. (2014). Native-likeness in second language lexical categorization reflects individual language history and linguistic community norms. *Front. Psychol.* 5:1203. doi: 10.3389/fpsyg.2014.01203

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.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 13 August 2014
doi: 10.3389/fpsyg.2014.00888

Phonological and orthographic cues enhance the processing of inflectional morphology. ERP evidence from L1 and L2 French

Haydee Carrasco-Ortiz [1] ***and Cheryl Frenck-Mestre*** [2,3] ******

[1] *Universidad Autónoma de Querétaro, Santiago de Querétaro, Mexico*
[2] *Laboratoire Parole et Langage, Centre National de Recherche Scientifique, Aix-en-Provence, France*
[3] *Aix-Marseille Université, Aix-en-Provence, France*

Edited by:
Christos Pliatsikas, University of Kent, UK

Reviewed by:
Joana Acha, Basque Centre on Cognition, Brain and Language, Spain
Phaedra Royle, Université de Montréal, Canada

****Correspondence:***
Cheryl Frenck-Mestre, Laboratoire Parole et Langage, Centre National de Recherche Scientifique, 5 rue Pasteur, Aix-en-Provence, France
e-mail: cheryl.frenck-mestre@univ-amu.fr

We report the results of two event-related potential (ERP) experiments in which Spanish learners of French and native French controls show graded sensitivity to verbal inflectional errors as a function of the presence of orthographic and/or phonological cues when reading silently in French. In both experiments, verbal agreement was manipulated in sentential context such that subject verb agreement was either correct, ill-formed and orally realized, involving both orthographic and phonological cues, or ill-formed and silent which involved only orthographic cues. The results of both experiments revealed more robust ERP responses to orally realized than to silent inflectional errors. This was true for L2 learners as well as native controls, although the effect in the learner group was reduced in comparison to the native group. In addition, the combined influence of phonological and orthographic cues led to the largest differences between syntactic/phonological conditions. Overall, the results suggest that the presence of phonological cues may enhance L2 readers' sensitivity to morphology but that such may appear in L2 processing only when sufficient proficiency is attained. Moreover, both orthographic and phonological cues are used when available.

Keywords: ERPs, verbal inflection, late bilinguals, phonological processing, sentence processing

INTRODUCTION

Can one reed a book? Indeed, whether or not one necessarily activates phonological representations when reading and accessing the meaning of a word is a long standing debate in reading research (McCusker et al., 1981; Morris and Folk, 2000; Harm and Seidenberg, 2004). Early theories assumed that phonological recoding was a secondary, slower, route to meaning as compared to "direct access" via orthographic codes alone (Paap and Noel, 1991). This view has been seriously challenged in the last 15 years. Today's debate lies not in the question of whether phonological information is retrieved but how and when, i.e., whether phonological information is retrieved pre-lexically or only once a stored lexical form has been activated on the basis of orthography, thus giving rise to stored phonological information (for recent reviews see Van Orden and Kloos, 2005; Hino et al., 2013).

The strongest evidence for phonological mediation during the processing of written words has been provided by research on single-word reading (Van Orden, 1987; Lukatela and Turvey, 1993; Jared et al., 1999). Seminal behavioral work in this area has shown that the activation of phonological information is both rapid and automatic, while perhaps lagging one beat behind that of orthography (Perfetti and Bell, 1991; Ferrand and Grainger, 1993). Said findings have since been replicated using event-related potentials (ERPs; Grainger et al., 2006).

Less attention has been paid to the role of phonological information during sentence processing; however, those studies that have examined this question have shown benefits. In both English and in French, a "preview" of the phonological information contained in the upcoming word in the sentence facilitates the reading of said word (Pollatsek et al., 1992; Rayner et al., 1995, 1998; Miellet and Sparrow, 2004; Ashby et al., 2006; but see Daneman and Reingold, 2000). The effect is not restricted to alphabetic languages; it has also been reported for logographic languages, such as Chinese (Pollatsek et al., 2000; Liu et al., 2002; Tsai et al., 2004), The finding that the prior activation of phonological information enhances reading performance, whichever the language under question, adds considerable weight to the hypothesis that such is part and parcel of the natural reading process (although see Van Orden and Kloos, 2005; Hino et al., 2013 for further discussions of these phonological effects).

The present study examined the question of phonological recoding, but from a syntactic viewpoint. We examined the impact of phonological cues on the processing of subject–verb agreement in sentential context. Moreover, we examined this question in both first (L1) and second language (L2) processing. Indeed, this issue is important not only for understanding the contribution of phonology to morphological processing in native readers but also in understanding whether phonology can be useful in learning verbal inflection in an L2. The language under question was French, which presents a particularly interesting case to study how phonological representations may influence how the orthographic code is processed.

Processing written French requires dealing with numerous morphological inflections that do not have overt phonological representations. Regular inflections of the present tense are an

illustrative example of this phonetic opacity. Morphological variations are not systematically represented in the phonological code; verbal inflections of the three singular persons for regular (first group) verbs are phonologically identical, despite the orthographic variation for the second singular person [e.g. *tu parles* /tu ɒaʁlə/ "you speak" is pronounced the same as *je/il parle* /ʒə/il ɒaʁlə "I/he eat(s)"]. Analogously, the third person plural inflected form is phonologically identical to that of the three singular persons regardless of orthographic variation across forms [e.g. *ils parlent* /il ɒaʁlə/ "they speak" is pronounced the same as *je/tu/il parle(s)* "I/you/he speak(s)"]. The absence of audible distinctions for different morphological forms in written French seems to account for the majority of errors made in verbal agreement production (Brissaud and Sandon, 1999; Negro and Chanquoy, 2000; Largy and Fayol, 2001; Chevrot et al., 2003). These studies have indeed shown that the presence of an audible morphological marker considerably reduces verbal agreement errors in children and skilled literate adults alike. Akin to the results obtained for the French language, both Dutch children and experienced adults produce more spelling errors on regularly inflected verb forms that do not include audible differences than for those that do (Frisson and Sandra, 2002; Sandra et al., 2004).

The effect of phonological mediation on inflected morphology can also be seen in comprehension, although debate indeed remains as to the systematic nature of this effect. In a self paced reading study, Brysbaert et al. (2000) examined whether orthographic information is necessary and/or sufficient to process homophonous verb forms and how the reading system deals with this type of silent morphological information. To do so, they looked at the respective contribution of orthographic and phonological cues on the processing of verbal tense in short sentence contexts. Results for native Dutch readers showed that orthographic cues alone sufficed to process the tense of the verb. These results were interpreted as evidence that the reading system is sensitive to orthographic information that is not represented at the phonological level. However, according to the authors, these findings could not be taken as evidence against the use of phonological cues in normal reading. Indeed, the simple presence of homophonic verb forms may have triggered the alternative use of orthographic information in order to disambiguate the verb. In line with the model forwarded by Harm and Seidenberg (2004), numerous empirical studies have shown that orthographic and phonological information can be activated in parallel during silent reading (cf. Van Orden and Kloos, 2005, for a review).

Studies that used ERPs to investigate the processing of inflectional morphology during written sentence processing have provided further understanding about the role of phonological cues in this process, in both native and non-native speakers (Osterhout et al., 2004, 2006; Frenck-Mestre et al., 2008). In a longitudinal study, Osterhout et al. (2004) suggested that L2 learners can quickly integrate morphosyntactic features with minimal instruction (e.g. 4 months), especially when features are present in the native language and morphological inflections involve phonological markers. Results showed that English L1–French L2 learners elicited an ERP effect in response to verbal agreement violations that presented phonological cues [e.g. *Tu adores*adorez* "You adore(s)"], whereas no variation in the ERP responses was observed in response to determiner-noun agreement errors, which were both largely absent from the native grammar and that involved silent morphemes [e.g. *Tu manges des hamburgers*hamburger* "You eat hamburger(s)"]. However, it remained unclear whether it was the similarity of grammatical features across L1–L2 alone, or this factor combined with the oral realization of morphemes that lead to faster learning and a more systematic brain response to violations.

To isolate the effects of phonological realization on morphosyntactic processing, subsequent studies manipulated the presence versus the absence of phonological cues to grammatical morphemes for shared grammatical features in L1 and L2 (Frenck-Mestre et al., 2008; McLaughlin et al., 2010). Frenck-Mestre et al. (2008) examined the impact of phonological realization on verbal agreement in both native and non-native speakers of French. Results showed that compared to grammatically correct sentences, verbal agreement errors involving orally realized morphemes elicited a P600 effect for both native French speakers and German L1–French L2 learners. In contrast, silent inflectional errors produced different ERP patterns across the two groups. French native speakers showed a smaller P600 effect in comparison to orally realized errors whereas for L2 learners no robust effects were found in response to silent errors. These authors concluded that the presence of phonological information facilitates the processing of verbal agreement errors in native speakers and enhances the learning rate of verbal inflection in L2 learners of French.

The effects of phonological realization on morphosyntactic learning were also examined longitudinally in non-native speakers by McLaughlin et al. (2010). ERPs were recorded in three consecutive sessions while English L1–French L2 learners read the same materials as those used by Frenck-Mestre et al. (2008). At the end of the third session, a subgroup of learners presented an N400 effect to verbal agreement violations. By contrast, the other subgroup of learners, who first elicited an N400 response to inflectional errors, showed a subsequent small P600 effect in this third session. Although ERP differences were not observed as a function of whether the inflectional errors were phonologically realized or silent, learners' acceptability judgments were indeed sensitive to the presence of phonological cues. According to the authors, the processing of orthographic cues may not have triggered the activation of phonological information in L2 learners in contrast to the results found for native speakers and more proficient learners (Frenck-Mestre et al., 2008). However, the presence of oral cues does nonetheless seem to have an effect on morphological learning in the early stages of acquisition as evidenced by the learners' behavioral responses.

The current set of studies further investigated the extent to which phonological cues impact upon the processing of inflectional morphology in silent reading. Thus far, behavioral and ERP studies have presented inconsistent results with respect to the specific contribution of phonological cues during morphosyntactic processing in both native and non-native language readers (Brysbaert et al., 2000; Frenck-Mestre et al., 2008; McLaughlin et al., 2010). Indeed, different levels of sensitivity to phonological cues have been suggested in the literature discussed above.

Thus, in two experiments, we aimed to determine the extent to which native and non-native readers are sensitive to the presence of phonological cues when processing inflectional morphology. To do so, we recorded ERPs from native French speakers as well as from Spanish L1–French L2 learners while they read sentences that contained subject–verb agreement errors which were either phonologically realized, i.e., when morphological information involved both orthographic and phonological cues, or silent, i.e., when morphological information involved orthographic cues that were not orally realized (cf. **Table 1**).

In a first experiment, we wished to determine whether the absence of an effect of phonological cues to verbal morphology in L2 learners reported by McLaughlin et al. (2010) was indeed attributable to the relatively low proficiency of these learners. If phonological recoding only comes into play at a more advanced level of L2 proficiency, we should find that more proficient L2 learners show the effect. To address this question, we examined L1 Spanish speakers who were immersed in the French language and who had several years of formal instruction in this language. Based on the findings reported by Frenck-Mestre et al. (2008), we predicted that these more advanced learners would show sensitivity both online and offline to verbal inflected violations and that such would vary as a function of the presence of phonological cues.

In a second experiment, we examined whether the presence of orthographic cues in addition to phonological ones may enhance processing of inflectional morphology. Brysbaert et al. (2000) found that the added presence of phonological cues did not allow participants to recover the tense of verbs faster, as compared to both orthographically and phonologically ambiguous forms. They concluded that readers can process verbal inflection as easily when silent as when orally realized. To address this issue, in our second experiment we manipulated verbal inflections such that orthographic variation was held constant across varying phonological conditions. Both a new control group of native French speakers and a new group of advanced Spanish–French bilinguals were tested.

The two groups of Spanish–French late bilinguals recruited for the present study were considerably more advanced than the English–French late L2 learners who participated in the McLaughlin et al. (2010) study. Importantly, all of our L2 participants were enrolled in university level classes conducted in French at a French university and had been immersed in the French language for roughly 6 months at the time of the study. There is clear evidence that direct classroom instruction on phonological form can improve learning rate of French grammar (Arteaga et al., 2003). Nonetheless, as highlighted by Muñoz (2008, 2014), we can assume in line with the data from numerous linguistic and psycholinguistic studies that immersion will exert an important influence on L2 proficiency. As concerns the ability to compute syntactic structures, late L2 learners who have been immersed in the L2 for several years show a preference for resolving ambiguous structures in a manner similar to native speakers, quite unlike less experienced L2 learners who show an influence of their L1 (Frenck-Mestre, 2002; see also Dussias, 2003). It is important to underline that such a result is not a product of any explicit training but simply the by-product of extended experience with the L2 in daily life. More recently, Pliatsikas and Marinis (2013) reported successful online, immediate processing of complex syntactic structures (filler-gap dependencies) in a group of L2 English speakers who had been immersed for several years in their L2 environment but not for L2 learners with less exposure. Again, no specific training on said structures was given; with more years of experience, the late L2 learners simply achieve a high level of automatized syntactic parsing. Note, these findings are at odds with the claim that late L2 learners make do with shallow parsing based on heuristics (Clahsen and Felser, 2006), but argue in favor of models which assume that provided sufficient experience with/exposure to a language adult learners will achieve "nativelike" syntactic parsing (Herschensohn, 2000; Hopp, 2010; Steinhauer, 2014).

There is also some electrophysiological evidence that adult learners benefit from "naturalistic" or implicit learning as opposed to explicit instruction when confronted with a new language. Morgan-Short et al. (2012) reported that adult learners who were given implicit instruction on an artificial grammar, somewhat akin to what happens in immersion, showed "more native like" cortical responses to violations of word order than did a group of learners with equal exposure but explicit instruction. The pattern of ERP results was nonetheless complex. Participants were tested twice, once when at a low level of proficiency in the artificial language and again, following more training when they had achieved a high level of proficiency. Morgan-Short et al. (2012) reported that at higher proficiency, the implicit learning group showed a "native-like" pattern, consisting of an anterior-negativity followed by a P600 response to word order violations, whereas the

Table 1 | Examples of the three sentence conditions (correct, incorrect and phonologically realized, incorrect and silent) for the six different verbal persons in French.

Sentence onset	Correct	Incorrect, phonologically realised	Incorrect, phonologically silent	Sentence end
Le matin	je mange /mɑ̃ʒə/	mangez /mɑ̃ʒe/	manges	du pain
	tu manges	mangez	mange	
	il/elle mange	mangez	manges	
	nous mangeons	mangent /mɑ̃ʒə/		
	vous mangez	manges /mɑ̃ʒə/		
	ils/elles mangent	mangeons /mɑ̃ʒõ/	manges	

explicit learning group showed only a P600. First, as noted by many and recently reiterated by Steinhauer (2014) and Tanner (2014), the presence of an early negativity (whether left lateralized or not) in response to syntactic processing difficulty is not systematic enough to be considered the hallmark of automatic syntactic parsing, whether in a native or late learned language (but see Molinaro et al., 2011). Perhaps more convincingly, at the lower level of proficiency the implicit learning group showed an N400 effect (albeit preceded by an early negativity, from 150 to 300 ms and extending to 700 ms), which then became a P600 (albeit preceded by the same early negativity – which may simply have been the remnant of the N400) at the second testing period. The explicit training group showed no ERP response when at a low level of proficiency. Hence, if nothing else, the implicit learning elicited a cortical response earlier than did explicit training. It is of interest that this is the first study, to our knowledge, to show an "N400–P600" transition for aurally presented sentences, and for artificial grammars.

In both experiments, we expected the cortical response to grammatical violations to vary as a function of the presence of overt phonetic cues to grammatical morphemes. Exactly what ERP signature we could expect to show variation depends upon the particular study that one considers. In Frenck-Mestre et al. (2008) variation in the P600 response was found in both native and L2 speakers but of different types; in native speakers, orally realized inflectional errors produced a larger P600 response than silent errors. In L2 speakers, orally realized errors elicited a significant P600 response while silent errors produced only a hint of variation and in the N400 response. Seminal work by Osterhout et al. (2004) has shown that adult L2 learners can show either the more typical P600 response or an N400 effect when confronted with violations of inflectional morphology, and that some L2 learners show a gradual shift from an N400 effect to a more typical P600 response as a function of grammatical competence. This result has since been replicated, both in late L2 learners (McLaughlin et al., 2010; Tanner et al., 2013) and in "heritage" L2 speakers, i.e., speakers whose parents are native speakers but who were raised and schooled in an environment where the ambient language was not that of their parents (Tanner et al., 2013). Other L2 studies have also reported N400 effects in response to grammatical violations, but as a function of grammatical structure rather than competence, with the same L2 speakers showing either an N400 or a P600 depending upon the familiarity with/frequency of structures in the L2 (Foucart and Frenck-Mestre, 2012). It is of importance to note that an N400 effect to syntactic violations is not specific to L2 processing. Osterhout (1997) clearly demonstrated that native speakers can show an N400 effect rather than a P600 in response to violations of syntactic expectancies, especially if the critical word is sentence final. This result, although often swept under the rug in L1 studies of syntactic processing where a P600 dominates in the averaged waveform, especially if the critical word is sentence medial, is well known to any who have looked at their individual data (see also Osterhout et al., 2004). This issue has recently been revisited by Tanner and Van Hell (2014). It has also been shown that the magnitude of the P600 response in native speakers is highly linked to their linguistic proficiency (Pakulak and Neville, 2010). Given all of the above, it is plausible to assume that we might find variation in either of these components as a function of the well formedness of our sentence materials and the overt phonological realization thereof. The debate about what these cortical signatures reveal about sentential processing will be presented in the general discussion and in light of the results obtained here (for discussions, see Steinhauer et al., 2009; Molinaro et al., 2011; Tanner, 2014).

EXPERIMENT 1

The goal of Experiment 1 was to examine the extent to which advanced Spanish L1–French L2 learners rely on phonological cues to process inflectional morphology online. In this aim, we examined the ERP responses of participants while they read sentences that contained either phonologically realized or silent subject–verb agreement errors. If phonological cues are indeed available online to advanced L2 learners, we should replicate previous results (Frenck-Mestre et al., 2008) showing that these cues increase L2 readers' sensitivity to morphological violations in like fashion to native controls. If phonological cues are not, however, taken into account as systematically during online L2 processing we can predict offline differences for silent and orally realized errors but not variations in the ERP response for the L2 group. Indeed, the question remains open whether the lack of an online effect of phonological cues in the L2 reported by McLaughlin et al. (2010) is attributable to the lower proficiency of the L2 learners as compared to those studied by Frenck-Mestre et al. (2008) or to a generally less systematic use of these cues during L2 processing. To ascertain whether the pattern for our late Spanish–French bilinguals was similar to that previously reported for native French speakers, we report data for a control group, which largely overlapped with that reported in Frenck-Mestre et al. (2008).

METHOD

Participants

Sixteen native Spanish speakers (eight female) aged 21–29 years (mean age 24.2 years) participated in the study (one was excluded from analyses due to excessive artifacts). All were classifiable as "late bilinguals" (mean age of acquisition of French, 16.8 years, and mean years of study of French, 4.7 years). All had passed the second level of the DELF, a standardized test of French as a second language, were following a university curriculum in the French language and were living in France (mean of 5.5 months) at the time of participation. Their mean self-rating of reading expertise in the French language (on a scale from 1 to 6) was 4.0 (SD = 0.9). Another group of 15 participants were native French speakers (seven female) aged 20–24 years (mean age 21.5 years), enrolled at a French university (one participant was replaced due to excessive movement during ERP recording; 13 of these participants were the same as reported in Frenck-Mestre et al., 2008). All participants – French and Spanish – had also learned English as a second language throughout secondary school, although their fluency in this language was not tested. All participants were dominant right handed with normal or corrected-to-normal vision. Participants were paid for their participation. They all signed an informed consent form for the study, which was approved by the French

ethics committee, and were fully debriefed at the end of the experiment.

Materials

Critical stimuli were twenty regular French verbs of the first group presented in 90 declarative present-tense sentences. Grammaticality of sentences was manipulated by verbal agreement between the subject pronoun and the verb. All six verbal persons were used. Three morphosyntactic conditions were created by manipulating the pairing of verbal person and verbal inflection, with 30 sentences per condition: correct (e.g. "je parle" /ʒə ˌpaʁlə/), incorrect and orally realized (e.g. "je parlez" /ʒə ˌpaʁle/), and incorrect and silent (e.g. "je parles" /ʒə ˌpaʁlə/). Examples are provided in **Table 1**. In the correct and phonologically realized incorrect conditions, the three singular verbal persons were seen an equal number of times (four times each) as were the 3 plural verbal persons (six times each), In the phonologically silent incorrect condition, the first person plural ("nous") and second person singular formal/plural ("vous") were not included as any mis-pairing gives rise to an overt phonological form; the other four forms (1st, 2nd informal and 3rd singular, 3rd plural) were seen equally. Sentences were from 5 to 10 words in length and critical verbs appeared at varying word positions, from the second to the fifth word, but never in the final position. Each verb was presented either four or five times across the 90 sentences. Forty-five additional filler sentences that did not involve morphosyntactic anomalies were included as filler sentences to distract participants' attention from the morphosyntactic manipulation. This fillers also included correct verbal inflections of the different verbal persons in such a way that all verbal persons were seen an equal number of times in correct and incorrect conditions across the entire set of materials. A latin square design was used, such that each experimental sentence was rotated across three lists, with each occurring only once per list and in a different condition per list. Each list contained 90 experimental sentences (30 per condition) and 45 filler sentences. Each participant saw only one list and three different random orders of presentation of sentences were created per list.

Procedure

Participants were instructed to read sentences silently from a computer monitor while seated comfortably in an isolated room. Each trial sequence consisted of the following: a fixation cross (500 ms) followed by the stimulus sentence, which was presented visually one word at a time, each word being displayed for 450 ms followed by a 150 ms blank-screen inter-stimulus interval, followed by a "oui/non" prompt. Participants read for comprehension and made meaning-acceptability judgments at the prompt after each sentence by means of a button box.

EEG activity was recorded continuously from 21 scalp locations referenced to the left mastoid, with a sampling rate of 200 Hz. Two additional electrodes were used to monitor for horizontal and vertical eye movements. Epochs began 100 ms prior to stimulus onset and continued 1100 ms thereafter. Average ERPs were calculated off-line from trials free of artifacts (less than 3% of rejections per condition overall; no differences in rejection rate were found as a function of condition or participant group). Averaging was performed without regard to behavioral responses[1].

Data analysis

Mean voltage amplitudes and peak latencies were calculated for two time windows: 400–550 and 600–800 ms. These time epochs have been associated with the N400 and/or anterior negativities and P600 components, respectively. Data acquired at midline and lateral sites were treated separately. A three-way ANOVA was performed on the mean amplitude and peak latency acquired at midline, with two levels of Group (Native French vs. Spanish–French bilinguals), and repeated measures on three levels of Verbal inflection (correct inflection, orally realized and silent inflectional errors) and Electrode (frontal, central, and parietal). Five-way ANOVAs were performed on the data acquired at lateral sites, involving Group (Native French vs. Spanish–French bilinguals), and repeated measures on three levels of Verbal inflection (correct inflection, orally realized and silent inflectional errors), two levels of Hemisphere (left, right), two levels of Site (anterior, central–parietal) and three levels of Electrode (three anterior and three central–parietal per hemisphere). The Greenhouse and Geisser (1959) correction was applied to all repeated measures with greater than one degree of freedom. All significant differences involving more than two conditions were confirmed by *post hoc* comparisons (Bonferroni).

RESULTS

Behavioral data

Grammatical acceptability judgments were analyzed as a function of Group (native French vs. Spanish L1–French L2 learners) and Verbal inflection (correct, orally realized, and silent errors). There was a main effect of Verbal inflection [$F(2,56) = 11.52$, $p < 0.01$] that tended to be modified by Group [$F(2,56) = 3.42$, $p < 0.06$]. *Post hoc* comparisons revealed the main effect of Verbal inflection to be due to the difference in correct responses for silent errors compared to orally realized errors and correct inflections ($p < 0.01$). This effect was further confirmed in each group independently [French ($F(2,28) = 10.23$, $p < 0.01$) and Spanish L1–French L2 learners ($F(2,28) = 3.39$, $p < 0.05$]. Thus, the difference in the percentage of correct responses for silent errors compared to orally realized errors and correct inflections was bigger in the native French speakers in comparison to that observed in Spanish L1–French L2 learners [mean percentage of correct detections for correct sentences, orally realized and silent errors were for native French speakers 93% (SD 3.1), 96% (SD 2.9), and 72% (SD 8.1), respectively, and for Spanish L1–French L2 learners 91% (SD 3.4), 89% (SD 5.6), and 82% (6.1), respectively].

[1] In both experiments, ERP data were analyzed based on all trials that were not contaminated by artifact. Response-contingent analyses are problematic in the present study given the unequal number of trials across experimental conditions. Indeed, in the phonologically silent condition, there was a substantial percentage of trials on which participants, whether native or L2 speakers, failed to consciously report an error, which was not the case for sentences in the orally realized condition. We nonetheless looked at the ERP data based on response-contingency. The results reported herein held up for all comparisons between orally realized inflectional errors and correct inflections. The results for the silent inflectional errors were less robust. Nonetheless, since excluding those trials in which the behavioral response was incorrect led to excluding more than 40% of the data, the reliability of these analyses is questionable.

Event-related potentials

Grand-average waveforms to the critical verbs for each verbal agreement condition are shown in **Figure 1** for native French and in **Figure 2** for Spanish L1–French L2 participants. As is visible in the figures, a clear "N1–P2" complex was evoked in the first 300 ms following critical word onset, for all conditions. Following this, in comparison to correctly inflected verbs, inflectional errors provoked a positive deflection that began at roughly 500 ms and persisted until around 800 ms, with a peak at roughly 600 ms, generally described as a P600 effect. This was true for both orally realized and silent errors and was observed for both native French speakers and for Spanish–French late bilinguals. In addition, a larger P600 effect was apparent for the orally realized errors compared to silent errors in both participant groups. Statistical comparisons confirmed these differences.

100–300 ms epoch

Statistical analyses revealed no reliable differences across conditions in the first 300 ms following critical verb onset.

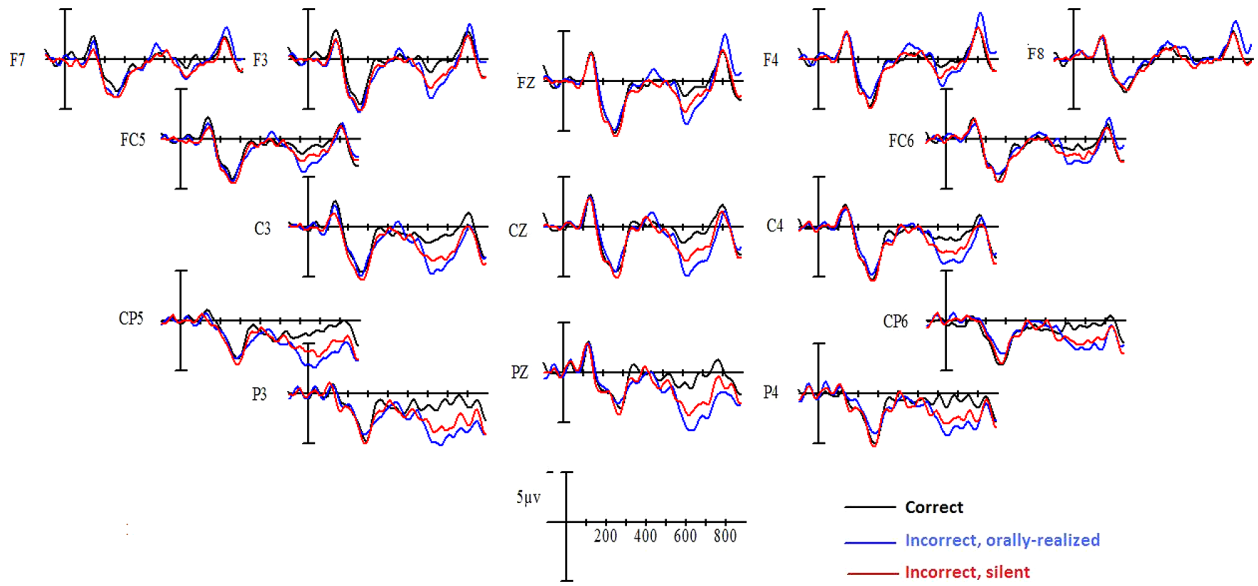


FIGURE 1 | Grand mean averages for native French speakers as a function of verbal inflection condition and electrode site.

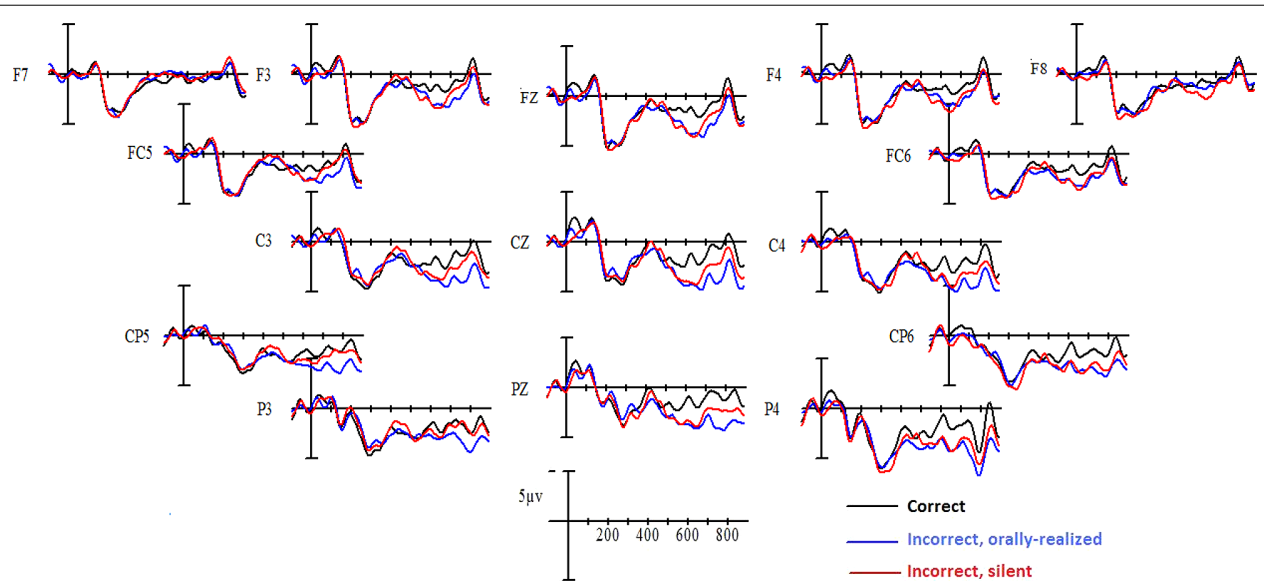


FIGURE 2 | Grand mean averages for Spanish L1–French L2 speakers as a function of verbal inflection condition and electrode site.

400–600 ms epoch

Analysis of data in this time window revealed no effects at midline. For lateral electrodes, an effect emerged for participant Group [$F(1,28) = 5.26$, $p < 0.02$] which was modified by interactions with Hemisphere × Verbal inflection [$F(2,56) = 3.08$, $p < 0.05$]. Independent analyses performed on the data for each participant group showed this interaction to be due to the presence of a positive deflection in response to both orally realized and silent errors only for Spanish L1–French L2 learners. No such early P600 effect was apparent in the native French group for whom no reliable differences as a function of experimental factors were found in this time window.

600–800 ms epoch

Analysis of data in this time window yielded a main effect of Verbal inflection [midline: $F(2,56) = 18.41$, $p < 0.0001$; lateral sites $F(2,56) = 9.67$, $p < 0.001$]. At midline, compared to well-formed sentences both orally realized and silent inflectional errors produced a significant P600 effect ($p < 0.01$ or better for both cases), which was significantly larger for orally realized errors than silent errors as confirmed by Bonferroni *post hoc* comparisons ($p < 0.01$). In addition, there was a trend for the interaction involving Verb, Group, and Electrode site [$F(4,112) = 2.29$, $p < 0.10$]. At lateral sites, there was a significant interaction between Verb, Group, and Hemisphere [$F(2,56) = 3.83$, $p < 0.03$]. Given such, independent analyses were performed on the data in this time window for each of the two participant groups.

In the native French group, an effect of Verbal inflection was observed at midline [$F(2,28) = 7.11$, $p < 0.003$] which was modified by Electrode [$F(4,56) = 8.44$, $p < 0.001$]. *Post hoc* comparisons revealed that compared to grammatically correct cases, orally realized errors elicited a significant P600 response at all three midline electrodes ($p < 0.01$ or better) whereas silent errors produced a reliable P600 response at central ($p < 0.001$) and parietal ($p < 0.001$) but not at frontal sites. In addition, orally realized errors produced a significantly larger P600 effect than silent errors at central ($p < 0.03$) and parietal sites ($p < 0.001$). The effect of Verbal inflection was also significant at lateral sites [$F(2,28) = 3.91$], revealing a significant P600 effect for both orally realized and silent errors. The P600 effect was present over central–parietal but not over anterior sites and it varied as a function of Verbal inflection [$F(2,28) = 8.57$, $p < 0.004$]. *Post hoc* comparisons revealed no P600 effect at anterior lateral sites, whereas at central–parietal lateral sites both types of error produced a robust P600 effect. In addition, this P600 effect was larger for orally realized errors ($p < 0.0001$) than for silent errors ($p < 0.001$), although the direct comparison of the two error conditions at posterior lateral sites did not show a reliable difference.

In the Spanish L1–French L2 learners group, a significant effect of Verbal inflection was observed at midline [$F(2,28) = 13.8$, $p < 0.0001$], which did not interact with Electrode. *Post hoc* comparisons at midline sites confirmed that compared to control sentences, orally realized and silent inflectional errors both produced a P600 effect ($p < 0.01$ or better) and differed from each other at all three midline sites ($p < 0.05$). At lateral sites, there was also an effect of Verbal inflection [$F(2,28) = 7.87$, $p < 0.01$], which tended to be modified by Hemisphere [$F(2,28) = 3.37$, $p < 0.06$] and was significantly so by the interaction involving Verbal inflection, Site (anterior/posterior), and Electrode [$F(4,56) = 10.45$, $p < 0.001$]. At anterior lateral sites, *post hoc* comparisons revealed that orally realized errors produced a more widespread P600 effect over the left than right hemispheres whereas the effect for silent errors was only significant over the right hemisphere ($p < 0.03$). At posterior lateral sites, an effect of Verbal inflection was observed [$F(2,28) = 10.83$, $p < 0.01$], with orally realized errors producing a robust widespread P600 effect ($p < 0.001$) and silent errors producing a more reduced effect ($p < 0.05$) that was also significantly smaller than that produced by orally realized errors over some electrodes ($p < 0.002$).

DISCUSSION

The results of Experiment 1 clearly demonstrate online sensitivity to morphosyntactic violations for Spanish L1–French L2 learners, in like manner to native French speakers, as revealed by a P600 response to these violations. Of principal interest in the present study was whether phonological cues would enhance the processing of morphosyntactic violations during silent reading. For native speakers, orally realized errors produced a greater P600 effect than silent errors, thus confirming previous results (Frenck-Mestre et al., 2008). In the same way, Spanish L1–French L2 learners showed a graded sensitivity to verbal inflection violations as a function of the presence of phonological cues. Indeed, the P600 response to orally realized errors was significantly larger than that observed to silent errors. The differentiated ERP response in non-native speakers contrasts nonetheless with recent work by McLaughlin et al. (2010). Using the same materials as in the present study but with less advanced English learners of French, these authors reported differences in sensitivity to verbal person violations as a function of whether these violations were orally realized or silent, but only on grammaticality judgments, not in the ERP response to these violations as reported here. In line with McLaughlin et al. and in view of our own results, we can forward the hypothesis that in less advanced L2 learners online processing may not be rapid and/or systematic enough to show a robust effect of phonological realization during reading in the ERP trace. We will re-examine this question in the general discussion in light of the results of Experiment 2.

To further investigate the role of phonology during silent reading and how such may impact grammatical processing as a function of L2 experience, we conducted a second experiment in which we again manipulated phonological variation. In line with interactive models of phonology and orthography during reading (Harm and Seidenberg, 2004) it is plausible that readers benefited from all cues available to them, whether phonological or orthographic. It is possible that the inclusion of plural pronouns in our first experiment may have enhanced the effect of oral realization that we observed, due to differences in orthography. Indeed, for the three singular pronouns the orthographic overlap between correct inflections and both the orally realized and silent violations was identical (one letter different in each case). Such was not the case for the plural, for which silent errors had more letters in common with correct inflections than orally

realized errors did (for a more indepth discussion, see Frenck-Mestre et al., 2010). To address this question and to ascertain that the differences in processing we observed were indeed due to the oral realization of morphology, we conducted a further experiment in which orthographic cues were reduced. This was done by restricting the verbal person manipulation to the three singular pronouns (1st, 2nd, and 3rd person). In said case, it is possible to hold orthographic variations constant meanwhile varying phonological overlap.

EXPERIMENT 2

In Experiment 1, we manipulated the presence vs. absence of oral cues to verbal agreement by mispairing inflection for both singular and plural pronouns. Our results clearly revealed an effect of phonological realization on agreement processing for both participant groups. The goal of Experiment 2 was to determine whether the effect of phonological realization observed in Experiment 1 was principally driven by phonological rather than orthographic cues. If so, we would expect this effect to persist when the amount of orthographic mismatch was identical across incorrect agreement conditions. Therefore, we used the three singular pronouns ("je," "tu," "il/elle"), for which verbal inflections vary by one letter and mispairing inflections may result in either orally realized or silent errors (see **Table 2**). We can predict that the presence of oral cues should enhance the reader's capacity to detect these errors. Based on previous ERP studies of written materials, we expected a variation in the amplitude of P600 and/or early negativities to verbal agreement errors as a function of oral cues, for native French and Spanish L1–French L2 speakers. It is also possible that an N400 effect could be elicited by morphological mispairings, especially in the L2 group. Indeed, various studies of adult L2 learners have reported N400 effects to just such errors, although none have manipulated the presence of overt oral cues to morphological variation (McLaughlin et al., 2010; Foucart and Frenck-Mestre, 2012; Morgan-Short et al., 2012; Tanner et al., 2013, 2014). The group of Spanish–French learners was, akin to the group tested in Experiment 1, immersed in French at the time of testing. All L2 participants had also had several years of formal study of the French language. As such, this group was comparable to that tested in the first experiment and considerably more advanced than the L2 learners tested by McLaughlin et al. (2010). Again, if L2 proficiency, as determined by either formal learning or amount of exposure is crucial to using phonological cues during reading in general and syntactic processing in particular, then we can predict similar results as found in Experiment 1, provided that the phonological cues are sufficient to produce effects in the absence of additional orthographic cues.

METHOD

Participants

Fifteen native French speakers (eight female) aged 18–25 years (mean age 20 years) and 15 native Spanish speakers (nine female) aged 21–29 (mean age 26.8 years) participated in this study. None of the participants had taken part in the first experiment. All subjects were dominant right handed with normal or corrected-to-normal vision. They were paid for their participation and signed an informed consent form for the study, which was approved by the French ethics committee. They were fully debriefed at the end of the experiment. Spanish speakers were classifiable as "late bilinguals" (mean age of acquisition of French, 23.6 years, and mean years of study of French, 3.5 years). All had passed the second level of the DELF, a standardized test of French as a second language, were following a university curriculum in the French language and were living in France (mean of 6 months) at the time of participation. Their mean self-rating of reading expertise in the French language (on a scale from 1 to 6) was 4.2 (SD = 0.77). All participants – French and Spanish – had also learned English as a second language throughout secondary school, although their fluency in this language was not tested.

Materials

An entirely new set of sentences was created for the purpose of this experiment. Critical stimuli were 30 regular French verbs from the first group (20 taken from Experiment 1) which were presented in 90 declarative present-tense sentences. Grammaticality of sentences was manipulated by verbal agreement between the subject pronoun and the verb. As in Experiment 1, 3 morphosyntactic conditions were created by manipulating the pairing of verbal person and verbal inflection, with 30 sentences per condition: correct (e.g. "je regarde"), incorrect and orally realized (e.g. "je regardez"), and incorrect and silent (e.g. "je regardes"). In contrast to Experiment 1, only the 3 singular persons were included. Sentences were from 5 to 10 words in length and critical verbs appeared at varying word positions, from the second to the fifth word, but never in the final position. Three lists were created such that each experimental sentence was rotated across lists in a Latin square design, with each occurring only once per list and in a different condition per list. In each condition, the three singular pronouns were seen an equal number of times (10 per each, with five masculine and five feminine for the third person singular). Sixty additional filler sentences

Table 2 | Examples of the three sentence conditions (correct, incorrect and orally realized, incorrect and silent) for the three singular verbal persons in French used in Experiment 2.

Sentence onset	Correct	Incorrect, phonologically realized	Incorrect, phonologically silent	Sentence end
Le soir	je regarde	regardez	regardes	des films
	tu regardes	regardez	regarde	
	il/elle regarde	regardez	regardes	

that did not involve morphosyntactic anomalies were included to distract participants' attention from the syntactic manipulation. These fillers also provided a balance for correct and mispaired verbal inflections across the entire set of materials. Each participant saw only one list and three different random orders of presentation of sentences were created per list.

Procedure

The procedure was identical to that of Experiment 1.

Data analysis

This was identical to that of Experimental 1 with the exception that the time epoch associated with the N400 and/or anterior negativity was examined in the 300–500 ms time window, and that associated with the P600 component was shifted to the 500–700 ms and 700–900 time windows based on visual inspection of waveforms.

RESULTS

Behavioral data

Behavioral responses were analyzed based on grammatical acceptability judgments. A repeated-measures ANOVA involving Group (Native French vs. Spanish–French bilinguals) as a between subjects factor and Verbal inflection (correct, silent, and orally realized errors) as a repeated measure revealed a main effect of Verbal inflection [$F(1,43) = 27.42$, $p < 0.01$] which did not interact with Group ($F < 1$). Orally realized errors were better detected than silent errors and this was true for both participant groups [mean percentage of correct detections for correct sentences, orally realized and silent errors were 94% (SD 2.3), 92% (SD 3.5), and 73% (SD 8.2) for native French speakers and 89% (SD 4.7), 86% (SD 6.4), and 61% (SD 15.9) for Spanish L1–French L2 learners, respectively]. Given the differences across conditions as concerns the detection of errors, ERPs were calculated independent of responses. Moreover, as shown by numerous studies, behavioral responses are not necessarily indicative of cortical sensitivity (McLaughlin et al., 2004; Foucart and Frenck-Mestre, 2012; but see Foucart and Frenck-Mestre, 2011).

Event-related potentials

Grand-average ERPs elicited by critical verbs in the three verbal agreement conditions (correctly inflected verbs, orally realized, and silent errors) are shown in **Figure 3** for native French and in **Figure 4** for Spanish L1–French L2 learners. Visual inspection of these waveforms revealed a clear "N1–P2" complex evoked in the first 300 ms following critical word onset, for all conditions. After 300 ms, two effects emerged. First, a negative component was observed over the left and right hemispheres, beginning around 300 ms and persisting until 500 ms, for inflectional errors as compared to correct verbal agreement. This negativity was observed predominantly in the native French participant group and mainly for orally realized inflectional errors. Following this negativity, inflectional errors provoked a positive deflection in comparison to correctly inflected verbs. This positivity presented different onset latencies across participant groups. For native French speakers positivity began at approximately 500 ms and persisted until roughly 800 ms, whereas for Spanish L1–French L2 learners the positivity began at 700 ms and persisted beyond 800 ms. Thus the P600 effect was delayed in L2 learners. ANOVAs performed on the mean amplitude data confirmed these effects.

300–500 ms epoch

Statistical analyses in this time window yielded a significant Verbal inflection by Group interaction at both midline [$F(2,56) = 3.79$, $p < 0.03$] and lateral sites [$F(2,56) = 3.70$, $p < 0.03$]. Independent analyses were subsequently performed on the data for each participant group. For native French speakers, a main effect of Verbal inflection was observed only at lateral sites [$F(2,28) = 4.06$, $p < 0.04$]. Planned comparisons confirmed reliable differences between correctly inflected verbs and orally realized errors at lateral sites ($p < 0.03$) while no differences were observed between correctly inflected verbs and silent errors ($p > 0.15$). The comparison of orally realized and silent errors showed a reliable difference ($p < 0.03$). For Spanish L1–French L2 learners, the main effect of Verbal inflection approached significance only at midline [$F(2,28) = 2.97$, $p < 0.08$]. *Post hoc* comparisons revealed that silent inflectional errors were more negative compared to correctly inflected verbs ($p < 0.02$). No differences between orally realized errors and correctly inflected verbs or between the two error conditions were found in this time window for this group.

500–700 ms epoch

Analyses of the data in this time window showed a significant interaction between Group and Verbal inflection at midline [$F(2,56) = 4.36$, $p < 0.01$], while at lateral sites Verbal inflection interacted with Group and Site [$F(2,56) = 2.65$, $p < 0.03$]. Separate analyses were subsequently performed on the data for each participant group. For native French speakers, a main effect of Verbal inflection was observed at midline [$F(2,28) = 4.35$, $p < 0.02$] and at lateral sites [$F(2,28) = 4.06$, $p < 0.02$]. *Post hoc* comparisons (Bonferroni) at midline showed that both orally realized and silent errors differed from correct sentences ($p < 0.01$), whereas orally realized and silent errors did not differ from each other. At lateral sites, the effect of Verbal inflection was modified by Site [$F(2,28) = 4.15$, $p < 0.05$]. Separate analyses at anterior and posterior sites revealed a main effect of Verbal inflection only at posterior sites [$F(2,28) = 4.25$, $p < 0.03$]. At posterior sites, *post hoc* comparisons confirmed reliable differences between correctly inflected verbs and orally realized errors ($p < 0.03$), while no differences were found between correctly inflected verbs and silent errors ($p > 0.15$). The direct comparison of the two error conditions showed a small trend ($p < 0.11$).

Analyses conducted on Spanish L1–French L2 speakers in this time window showed no effect of Verbal inflection or significant interactions at any electrode site ($F < 1$).

700–900 ms epoch

Analyses of data in this time window yielded a significant interaction between Verbal inflection and Group at midline [$F(2,56) = 6.27$, $p < 0.01$] and at lateral sites [$F(2,56) = 5.35$, $p < 0.01$]. Separate ANOVAs were subsequently conducted on the data for each participant Group. Native French speakers showed no significant effects or interactions at any electrode site ($F < 1$). In contrast, Spanish L1–French L2 speakers revealed a significant effect of Verbal inflection at midline [$F(2,28) = 5.93$, $p < 0.01$]

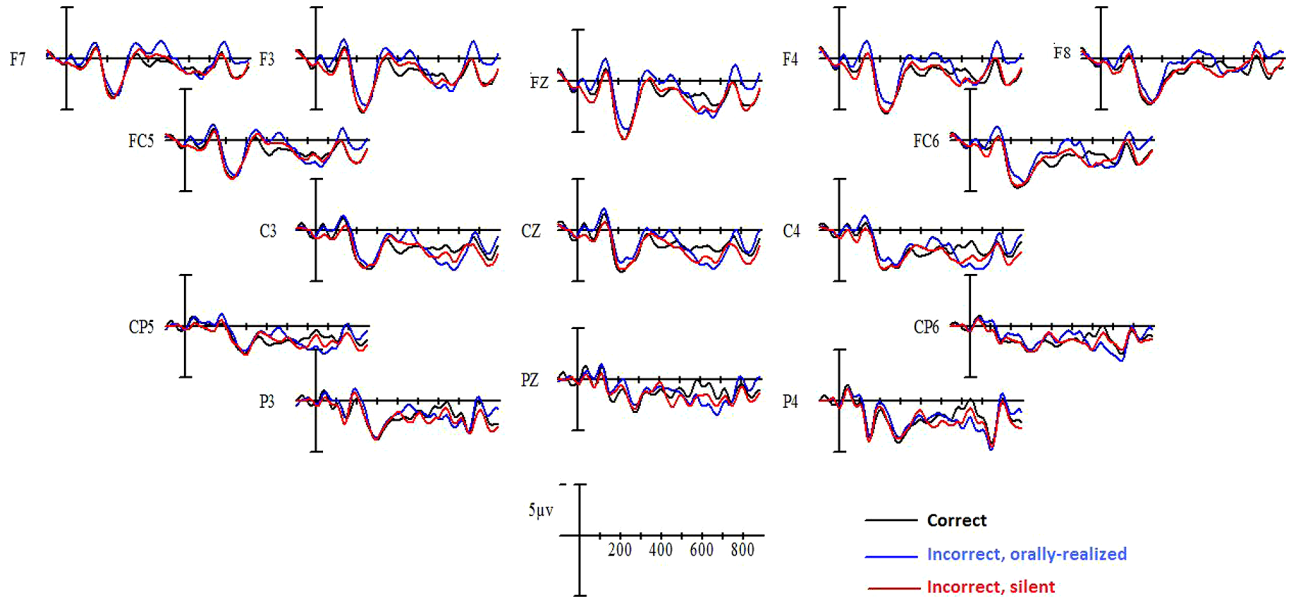


FIGURE 3 | Grand mean averages for native French speakers as a function of verbal inflection condition and electrode site.

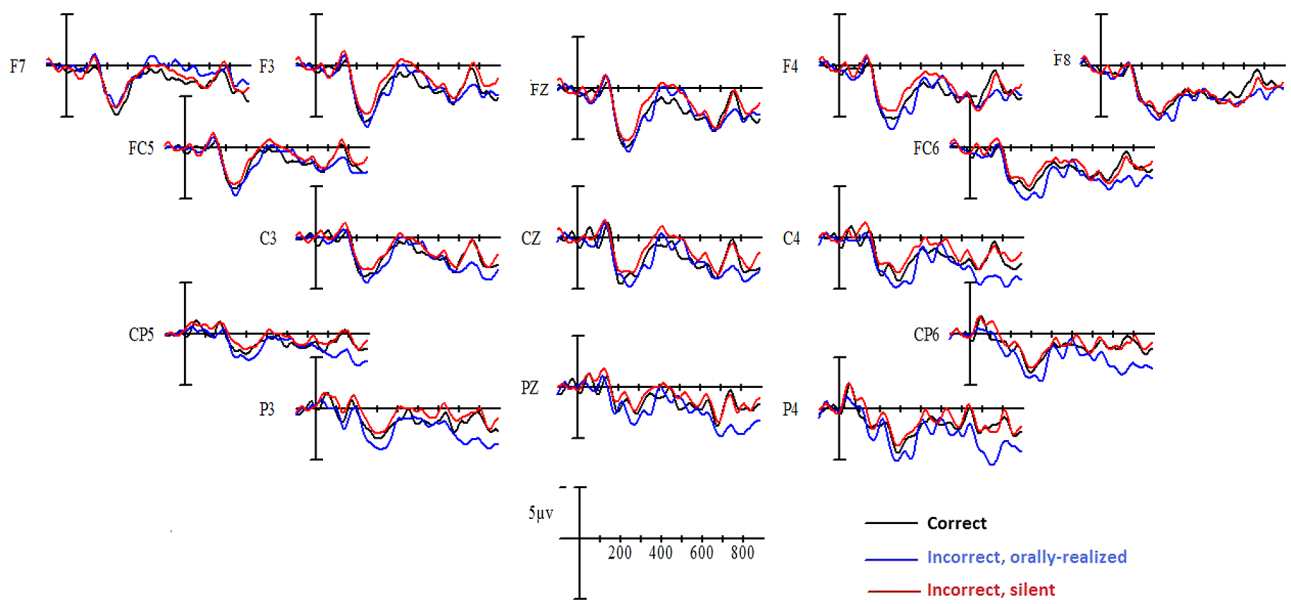


FIGURE 4 | Grand mean averages for Spanish L1–French L2 speakers as a function of verbal inflection condition and electrode site.

which was modified by Electrode site [$F(4,56) = 3.73$, $p < 0.03$]. At midline, *post hoc* comparisons revealed that compared to correct cases, orally realized errors elicited a significant P600 response at central ($p < 0.02$) and parietal ($p < 0.001$) sites, while no significant effect was observed at any electrode sites for silent errors, for any comparison. At lateral sites, a significant effect of Verbal inflection was observed [$F(2,28) = 4.27$, $p < 0.02$] which interacted with Site [$F(2,28) = 9.6$, $p < 0.01$]. *Post hoc* comparisons at anterior and posterior sites revealed that at posterior sites, orally realized but not silent errors elicited a significant P600 ($p < 0.01$) in comparison to correct sentences, whereas at anterior sites no reliable P600 response was observed. Further, orally realized errors produced a significant P600 effect compared to silent errors ($p < 0.01$).

DISCUSSION

The goal of Experiment 2 was to confirm that the effect of phonological realization observed in Experiment 1 could be attributed to the additional presence of phonological cues rather than to

orthographic cues alone. The results, obtained under conditions for which the amount of orthographic mismatch across verbal agreement conditions was controlled, showed again that orally realized verbal person violations provoked a reliable P600 response in comparison to correct cases and this was true for both native and non-native participants. In addition, compared to silent errors orally realized violations elicited a greater early negativity in native French speakers. Given its distribution, this effect can best be characterized as a member of the broad class of early negativities which has been reported in various studies (Rodriguez-Fornells et al., 2001; Hahne et al., 2006; Morgan-Short et al., 2012). In contrast to orally realized inflectional errors, silent errors elicited only a small effect. In the group of native French participants, silent errors only produced a significant P600 response at midline sites. For Spanish L1–French L2 speakers, the difference between orally realized and silent inflectional errors was clear and widespread; while orally realized errors elicited a P600, albeit in a late time window, silent errors did not produce any reliable effect in the ERP waveform.

GENERAL DISCUSSION

The present set of studies showed that phonological cues enhance the processing of inflectional morphology when reading silently in either one's first or second language. Specifically, the processing of verbal inflectional errors produced a larger P600 effect when violations involved both orthographic and phonological cues relative to when only inaudible morphological cues are available. These results provide evidence of the contribution of phonology to morphological processing in both native and non-native readers. The general findings here are consistent with those of previous studies that show an impact of phonological cues on the processing of inflectional morphology in French (Frenck-Mestre et al., 2008, 2010; Carrasco and Frenck-Mestre, 2009; McLaughlin et al., 2010) and confirm a systematic benefit from the presence of phonological cues under conditions where orthographic overlap across experimental conditions was held constant (Experiment 2). Although this was true for both native French speakers and Spanish L1–French L2 speakers, distinct neural responses were observed for each group as a function of the presence vs. absence of phonological cues when processing morphosyntactic markers.

For native speakers, a bilateral early negativity was evoked in response to orally realized errors in comparison to correctly inflected verbs. This was found, however, only when minimal orthographic differences between correct and orally realized errors were present (i.e., only in Experiment 2). In contrast, silent errors did not elicit any significant negativity, under any conditions. Both in terms of timing and distribution, the anterior negativity observed for orally realized errors fall within the range of variations that have been reported in previous studies for syntactic violations (Rodriguez-Fornells et al., 2001; Hahne et al., 2006; cf. Molinaro et al., 2011 for a review). The nature of this early negativity has been associated with a fast and automatic syntactic analyzer involving an initial detection of the grammatical error (Friederici, 1995, 2002; Hahne and Friederici, 1999). In line with this assumption, the fact that this negativity was present for orally realized but not for silent errors would suggest that phonological information has an effect on the first morphosyntactic analysis allowing a fast detection of orally realized inflectional errors. However, the interpretation of this early negativity should be considered with caution due to its lack of consistency across Experiments 1 and 2 and other previous studies (Frenck-Mestre et al., 2008, 2010; McLaughlin et al., 2010; Steinhauer, 2014; Tanner, 2014). Indeed, there is still considerable debate as to the significance and very nature of early negativities associated with syntactic processing in the monolingual literature (Osterhout, 1997; Molinaro et al., 2011; Tanner, 2014) which has seen repercussions in the literature on L2 processing (Frenck-Mestre, 2005; Steinhauer et al., 2009; Foucart and Frenck-Mestre, 2012; Steinhauer, 2014).

The results for native speakers showed that compared to orally realized errors, silent errors produced a smaller P600 effect in Experiment 1 and a reduced distribution in Experiment 2. The ERP differences observed for these two types of errors suggest that the presence of oral cues enhanced the syntactic analysis/reanalysis of violations of inflectional morphology in French. These results are in line with those obtained in previous off-line studies for native speakers of French (Negro and Chanquoy, 2000; Largy and Fayol, 2001) where phonologically realized morphemes induced fewer inflectional errors for verbal and nominal agreement in a written production task. It is noteworthy, nonetheless, that the difference observed between orally realized and silent inflectional errors was more pronounced in Experiment 1 in which the extent of orthographic differences between correctly inflected and erroneous cases was larger for orally realized than for silent errors. As such, the present results support the hypothesis that the added presence of orthographic cues indeed enhanced the effect of the oral realization of errors.

For the Spanish L1–French L2 speakers, no reliable early negativity was observed in response to verbal agreement errors. This result could fit, in a first instance, with the assumption that early negativities are restricted to native processing (Hahne, 2001) and or that they are associated with more advanced levels of processing (Steinhauer et al., 2009). However, the fact that various studies involving native speakers do not report any LAN effects in response to syntactic violations (Hagoort et al., 1993; Osterhout and Mobley, 1995; Osterhout et al., 2002; Frenck-Mestre et al., 2008; Foucart and Frenck-Mestre, 2011, 2012) renders difficult the interpretation of this absence of significant negativity effects in non-native speakers as a non "nativelike" processing. Indeed, the inconsistent presence of these LAN effects in native brain responses requires further research to reveal the underlying cognitive processes and the antecedent conditions that elicit or modulate it (Osterhout, 1997; Molinaro et al., 2011; Steinhauer, 2014; Tanner, 2014). Moreover, it has recently been suggested that individual differences in native speakers can account for the presence versus absence of early negativities to syntactic manipulations (Tanner, 2014). Clearly, further work is in order to clarify this issue.

In the present study, L2 learners showed a larger P600 effect when processing orally realized errors, as compared to silent errors and correctly inflected verbs. These results contrast with those reported in McLaughlin et al. (2010) where L2 learners showed no ERP difference between orally realized and silent errors. The greater on-line capacity to detect orally realized errors observed

in the present study might be associated with the relative high language proficiency in our L2 learners. Increased second language processing might enable the activation of phonological information, which could have enhanced the L2 learner's capacity to detect morphological violations involving phonological cues. In addition, the results of Experiment 2 replicate those reported in Frenck-Mestre et al. (2008), showing an absence of an ERP response to silent errors for L2 learners. One interpretation for these results is that non-native participants in Experiment 2 were not systematically sensitive to morphological errors that are not overtly realized. It is, therefore, possible that the L2 participants' response to silent inflectional errors was not strong enough to elicit a visible ERP response in the second experiment in contrast to the results of Experiment 1. This inherent heterogeneity of response in the L2 participants may have contributed to the absence of an effect for silent inflectional errors. Indeed, behavioral data in Experiment 2 showed that orally realized errors were detected significantly better than silent errors.

As outlined above, the effect of the phonological realization of morphology differed across experiments. Indeed, the difference in the ERP response to orally realized and silent errors was larger and more widespread for both participant groups in Experiment 1 than in Experiment 2. One possible explanation is that more robust effects are observed for experimental conditions in which the number of orthographic cues present in verbal inflection errors is greater. In Experiment 1, orally realized errors included all six verbal persons producing, therefore, orthographically more salient errors compared to Experiment 2 in which only the three singular persons were used (e.g., "$nous_{1st,\ plural}$ $parlent^{*}{}_{3rd,\ plural}$/$parlons_{1st,\ plural}$" compared to ("$je_{1st,\ sing}$ $parlez^{*}{}_{2nd,\ sing\ formal/plural}$/$parle_{1st,\ sing}$"). This relative orthographic advantage may have improved the processing of verbal inflection errors in Experiment 1. This finding is in line with assumption that readers benefit from all linguistic input when processing language (Brysbaert et al., 2000; Harm and Seidenberg, 2004; Frenck-Mestre et al., 2010). Under a connectionist framework, a mutual dependence of orthographic and phonological codes operates in the computation of a written word (Harm and Seidenberg, 2004). Thus, the different contribution of orthographic and phonological codes across experiments may have impacted the processing of words.

An important issue for this paper was to confirm the contribution of phonological cues to the online processing of inflectional morphology. In line with Brysbaert et al. (2000), the results obtained in the present study suggest that orthographic cues that are not phonologically represented in inflectional morphology can still be processed, though perhaps in a more effortful way. Our results also suggest that the presence of phonological cues can enhance the processing of inflectional morphology, even under conditions for which the amount of orthographic mismatch was identical. Furthermore, the present study provides important evidence relative to both native and non-native speakers' use of these cues during silent reading.

Finally, the impact of phonological cues on morphological variations is not limited to verbal processing. Indeed, the processing of other morphological inflections such nominal gender concord has been found to be enhanced by the presence of phonological cues (Carrasco and Frenck-Mestre, 2009; Foucart and Frenck-Mestre, 2011, 2012; for a review see Frenck-Mestre et al., 2010). In addition, this effect of phonology has been observed in both native and non-native speakers with diverse language backgrounds. Thus, the current set of experiments points to an active use of phonological information when reading silently and such is true, moreover in both first and second language processing.

REFERENCES

Arteaga, D., Herschensohn, J., and Gess, R. (2003). Focusing on phonology to teach morphological form in French. *Mod. Lang. J.* 87, 58–70. doi: 10.1111/1540-4781.00178

Ashby, J., Treiman, R., Kessler, B., and Rayner, K. (2006). Vowel processing during silent reading: evidence from eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 416–424. doi: 10.1037/0278-7393.32.2.416

Brissaud, C., and Sandon, J. M. (1999). L'acquisition des formes verbales en /E/à l'école élémentaire et au collège, entre phonographie et morphographie. *Lang. Franç.* 124, 40–57. doi: 10.3406/lfr.1999.6305

Brysbaert, M., Grondelaers, S., and Ratinckx, E. (2000). Sentence reading: do we make use of orthographic cues in homophones? *Acta Psychol.* 105, 31–56. doi: 10.1016/S0001-6918(00)00047-0

Carrasco, C., and Frenck-Mestre, C. (2009). "Phonology helps in processing grammatical gender: ERP evidence from L1 and L2 French," in *22nd Annual CUNY Conference on Human Sentence Processing*, Davis, USA, 26–28 March.

Chevrot, J. P., Brissaud, C., and Lefrançois, P. (2003). Norme et variations dans l'acquisition de la morphographie verbale en/E/: tendances, conflits de tendances, résolution. *Faits Lang.* 22, 57–66.

Clahsen, H., and Felser, C. (2006). How native-like is non-native language processing? *Trends Cogn. Sci.* 10, 564–570. doi: 10.1016/j.tics.2006.10.002

Daneman, M., and Reingold, E. M. (2000). "Do readers use phonological codes to activate word meanings? Evidence from eye movements," in *Reading as a Perceptual Process*, eds A. Kennedy, R. Radach, J. Pynte, and D. Heller (Oxford: Elsevier).

Dussias, P. E. (2003). Syntactic ambiguity resolution in L2 learners: some effects of bilinguality on L1 and L2 processing strategies. *Stud. Second Lang. Acquisit.* 25, 529–557. doi: 10.1017/S0272263103000238

Ferrand, L., and Grainger, J. (1993). The time course of orthographic and phonological code activation in the early phases of visual word recognition. *Bull. Psychon. Soc.* 31, 119–122. doi: 10.3758/BF03334157

Foucart, A., and Frenck-Mestre, C. (2011). Grammatical gender processing in L2: electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Biling. Lang. Cogn.* 14, 379–399. doi: 10.1017/S136672891000012X

Foucart, A., and Frenck-Mestre, C. (2012). Can late leaners aquire new grammatical features? Evidence from ERPs and eye-tracking. *J. Mem. Lang.* 66, 226–248. doi: 10.1016/j.jml.2011.07.007

Frenck-Mestre, C. (2002). "An on-line look at sentence processing in the second language," in *Syntactic Processing in the Second Language*, eds J. Altarriba and R. Herridia (North Holland: Elsevier).

Frenck-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language. *Second Lang. Res.* 21, 175–198. doi: 10.1191/0267658305sr257oa

Frenck-Mestre, C., Carrasco-Ortiz, H., McLaughlin, J., Osterhout, L., and Foucart, A. (2010). Linguistic input factors in native and L2 processing of inflectional morphology. Evidence from ERPs and behavioral studies. *Lang. Interact. Acquisit.* 2, 206–228. doi: 10.1075/lia.1.2.04fre

Frenck-Mestre, C., Osterhout, L., McLaughlin, J., and Foucart, A. (2008). The effect of phonological realization of inflectional morphology on verbal agreement in French: evidence from ERPs. *Acta Psychol.* 128, 528–536. doi: 10.1016/j.actpsy.2007.12.007

Friederici, A. D. (1995). The time course of syntactic activation during language processing: a model based on neuropsychological and neurophysiological data. *Brain Lang.* 50, 259–281. doi: 10.1006/brln.1995.1048

Friederici, A. D. (2002).Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* 6, 78–84. doi: 10.1016/S1364-6613(00)01839-8

Frisson, S., and Sandra, D. (2002). Homophonic forms of regularly inflected verbs have their own orthographic representations: a development

perspective on spelling errors. *Brain Lang.* 81, 545–554. doi: 10.1006/brln.2001.2546

Grainger, J., Kiyonaga, K., and Holcomb, P. (2006). The time course of orthographic and phonological code activation. *Psychol. Sci.* 17, 1021–1026. doi: 10.1111/j.1467-9280.2006.01821.x

Greenhouse, S. W., and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika* 24, 95–112. doi: 10.1007/BF02289823

Hagoort, P., Brown, C. M., and Groothusen, J. (1993). The Syntactic Positive Shift (SPS) as an ERP measure of syntactic processing. *Lang. Cogn. Process.* 8, 439–483. doi: 10.1080/01690969308407585

Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *J. Cogn. Neurosci.* 11, 193–204. doi: 10.1023/A:1010490917575

Hahne, A., and Friederici, A. (1999). Electropysiological evidence for two steps in syntactic analysis: early automatic and late controlled processes. *J. Cogn. Neurosci.* 11, 194–205. doi: 10.1162/089892999563328

Hahne, A., Mueller, J., and Clahsen, H. (2006). Morphological processing in a second language: behavioural and event-related potential evidence for storage and decomposition. *J. Cogn. Neurosci.* 18, 121–134. doi: 10.1162/089892906775250067

Harm, M. W., and Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychol. Rev.* 111, 662–720. doi: 10.1037/0033-295X.111.3.662

Herschensohn, J. (2000). *The Second Time Around: Minimalism and L2 Acquisition.* Philadelphia/Amsterdam: John Benjamins.

Hino, Y., Kusunose, Y., Lupker, S. J., and Jared, D. (2013). The processing advantage and disadvantage for homophones in lexical decision tasks. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 529–551. doi: 10.1037/a0029122

Hopp, H. (2010). Ultimate attainment in L2 inflectional morphology: performance similarities between non-native and native speakers. *Lingua* 120, 901–931. doi: 10.1016/j.lingua.2009.06.004

Jared, D., Levy, B. A., and Rayner, K. (1999). The role of phonology in the activation of word meanings during reading: evidence from proofreading and eye movements. *J. Exp. Psychol. Gen.* 128, 219–264. doi: 10.1037/0096-3445.128.3.219

Largy, P., and Fayol, M. (2001). Oral cues improve subject–verb agreement in written French. *Int. J. Psychol.* 36, 121–131. doi: 10.1080/00207590143000009

Liu, W., Inhoff, A. W., Ye, Y., and Wu, C. (2002). Use of parafoveally visible characters during the reading of Chinese sentences. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 1213–1227. doi: 10.1037/0096-1523.28.5.1213

Lukatela, G., and Turvey, M. T. (1993). Similar attentional, frequency, and associative effects for pseudohomophones and words. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 166–178. doi: 10.1037/0096-1523.19.1.166

McCusker, L. X., Gough, P. B., and Bias, R. G. (1981). Word recognition inside out and outside in. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 538–551. doi: 10.1037/0096-1523.7.3.538

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. doi: 10.1038/nn1264

McLaughlin, J., Tanner, D., Pitkanen, I., Frenck-Mestre,C., Inoue, K., Valentine, G., et al. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Lang. Learn.* 60, 123–150. doi: 10.1111/j.1467-9922.2010.00604.x

Miellet, S., and Sparrow, L. (2004). Phonological codes are assembled before word fixation: evidence from boundary paradigm in sentence reading. *Brain Lang.* 90, 299–310. doi: 10.1016/S0093-934X(03)00442-5

Molinaro, N., Barber, H., and Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex* 47, 908–930. doi: 10.1016/j.cortex.2011.02.019

Morgan-Short, K., Steinhauer, K., Stanz, C., and Ullman, M. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *J. Cogn. Neurosci.* 24, 933–947. doi: 10.1162/jocn_a_00119

Morris, R., and Folk, J. (2000). "Phonology is used to access word meaning during silent reading: evidence from lexical ambiguity resolution," in *Reading as a Perceptual Process*, eds A. Kennedy, D. Heller, and J. Pynte (Oxford: Elsevier), 427–446.

Muñoz, A. (2008). Age-related differences in foreign language learning. Revising the empirical evidence. *Int. Rev. Appl. Ling. Languist. Teach.* 46, 197–220.

Muñoz, A. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Appl. Ling.* doi: 10.1093/applin/amu024 [Epub ahead of print].

Negro, I., and Chanquoy, L. (2000). Subject–verb agreement with present and imperfect tenses: a developmental study from 2nd to 7th grade. *Eur. J. Psychol. Educ.* 15, 113–134. doi: 10.1007/BF03173170

Osterhout, L. (1997). On the brain response to syntactic anomalies: manipulations of word position and word class reveal individual differences. *Brain Lang.* 59, 494–522. doi: 10.1006/brln.1997.1793

Osterhout, L., McLaughlin, J., Allen, M., and Inoue, K. (2002). Brain potentials elicited by prose-embedded linguistic anomalies. *Mem. Cognit.* 30, 1304–1312. doi: 10.3758/BF03213412

Osterhout, L., McLaughlin, J., Kim, A., Greenwald, R., and Inoue, K. (2004)."Sentences in the brain: event-related potentials as real-time reflections of sentence comprehension and language learning," in *The On-line Study of Sentence Comprehension: Eyetracking, ERP, and Beyond*, eds M. Carreiras and C. Clifton Jr. (Brighton: Psychology Press).

Osterhout, L., McLaughlin, J., Pitkanen, I., Frenck-Mestre, C., and Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: a paradigm for exploring the neurocognition of second-language processing. *Lang. Learn.* 56, 199–230. doi: 10.1111/j.1467-9922.2006.00361.x

Osterhout, L., and Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *J. Mem. Lang.* 34, 739–773. doi: 10.1006/jmla.1995.1033

Paap, K. R., and Noel, R. W. (1991). Dual-route models of print to sound: still a good horse race. *Psychol. Res.* 53, 13–24. doi: 10.1007/BF00867328

Pakulak, E., and Neville, H. J. (2010). Proficiency differences in syntactic processing of native speakers indexed by event-related potentials. *J. Cogn. Neurosci.* 23, 2752–2765. doi: 10.1162/jocn.2010.21586

Perfetti, C. A., and Bell, L. (1991). Phonemic activation during the first 40 ms of word identification: evidence from backward masking and masked priming. *J. Mem. Lang.* 30, 473–485. doi: 10.1016/0749-596X(91)90017-E

Pliatsikas, C., and Marinis, T. (2013). Processing empty categories in a second language: when naturalistic exposure fills the (intermediate) gap. *Bilingualism* 16, 167–182. doi: 10.1017/S136672891200017X

Pollatsek, A., Lesch, M. F., Morris, R. K., and Rayner, K. (1992). Phonological codes are used in integrating information across saccades in word identification and reading. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 148–162. doi: 10.1037/0096-1523.18.1.148

Pollatsek, A., Tan, L. H., and Rayner, K. (2000). The role of phonological codes in integrating information across saccadic eye movements in Chinese character identification. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 607–633. doi: 10.1037/0096-1523.26.2.607

Rayner, K., Pollatsek, A., and Binder, K. S. (1998). Phonological codes and eye movements in reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 476–497. doi: 10.1037/0278-7393.24.2.476

Rayner, K., Sereno, S. C., Lesch, M. F., and Pollatsek, A. (1995). Phonological codes are automatically activated during reading: evidence from an eye movement priming paradigm. *Psychol. Sci.* 6, 26–32. doi: 10.1111/j.1467-9280.1995.tb00300.x

Rodriguez-Fornells, A., Clahsen, H., Lleó, C., Zaake, W., and Münte, T. (2001). Event related brain responses to morphological violations in Catalan. *Brain Res. Cogn. Brain Res.* 11, 47–58. doi: 10.1016/S0926-6410(00)00063-X

Sandra, D., Frisson, S., and Daems, F. (2004). Still errors after all those years: limited attentional resources and homophone frequency account for spelling errors on silent verb suffixes in Dutch. *Writ. Lang. Lit.* 7, 61–77. doi: 10.1075/wll.7.1.07san

Steinhauer, K. (2014). Event-related potentials (ERPs) in second language research: a brief introduction to the technique, a selected review, and an invitation to reconsider critical periods in L2. *Appl. Ling.* doi: 10.1093/applin/amu028 [Epub ahead of print].

Steinhauer, K., White, E., and Drury, J. E. (2009). Temporal dynamics of late second language acquisition: evidence from event-related brain potentials. *Second Lang. Res.* 25, 13–41. doi: 10.1177/0267658308098995

Tanner, D. (2014). On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement. *Cortex* doi: 10.1016/j.cortex.2014.04.007 [Epub ahead of print].

Tanner, D., Inoue, K., and Osterhout, L. (2014). Brain-based individual differences in on-line L2 grammatical comprehension. *Bilingualism* 17, 277–293. doi: 10.1017/S1366728913000370

Tanner, D., Mclaughlin, J., Herschensohn, J., and Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Biling. Lang. Cogn.* 16, 367–382. doi: 10.1017/S1366728912000302

Tanner, D., and Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia* 56, 289–301. doi: 10.1016/j.neuropsychologia.2014.02.002

Tsai, J.-L., Lee, C.-Y., Tzeng, O. J.-L., Hung, D. L., and Yen, N.-S. (2004). Use of phonological codes for Chinese characters: evidence from processing of parafoveal preview when reading sentences. *Brain Lang.* 91, 235–244. doi: 10.1016/j.bandl.2004.02.005

Van Orden, G. C. (1987). A ROWS is a ROSE: spelling, sound, and reading. *Mem. Cognit.* 15, 181–198. doi: 10.3758/BF03197716

Van Orden, G. C., and Kloos, H. (2005). *The Question of Phonology and Reading. The Science of Reading: A Handbook.* Oxford: Blackwell Publishing, 61–78. doi: 10.1002/9780470757642.ch4

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: 04 June 2014; accepted: 25 July 2014; published online: 13 August 2014.
Citation: Carrasco-Ortiz H and Frenck-Mestre C (2014) Phonological and orthographic cues enhance the processing of inflectional morphology. ERP evidence from L1 and L2 French. Front. Psychol. ***5****:888. doi: 10.3389/fpsyg.2014.00888*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 25 September 2014
doi: 10.3389/fpsyg.2014.01072

An ERP study on L2 syntax processing: When do learners fail?

Nienke Meulman[1]*, Laurie A. Stowe[1], Simone A. Sprenger[1], Moniek Bresser[2] and Monika S. Schmid[1,3]

[1] *Center for Language and Cognition, University of Groningen, Groningen, Netherlands*
[2] *Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Groningen, Netherlands*
[3] *Department of Language and Linguistics, University of Essex, Colchester, UK*

Edited by:
Christos Pliatsikas, University of Kent, UK

Reviewed by:
Christos Pliatsikas, University of Kent, UK
Eleonora Rossi, Penn State University, USA

****Correspondence:***
Nienke Meulman, Center for Language and Cognition, University of Groningen, Oude Kijk in 't Jatstraat 26, PO Box 716, 9700 AS Groningen, Netherlands
e-mail: n.meulman@rug.nl

Event-related brain potentials (ERPs) can reveal online processing differences between native speakers and second language (L2) learners during language comprehension. Using the P600 as a measure of native-likeness, we investigated processing of grammatical gender agreement in highly proficient immersed Romance L2 learners of Dutch. We demonstrate that these late learners consistently fail to show native-like sensitivity to gender violations. This appears to be due to a combination of differences from the gender marking in their L1 and the relatively opaque Dutch gender system. We find that L2 use predicts the effect magnitude of non-finite verb violations, a relatively regular and transparent construction, but not that of gender agreement violations. There were no effects of age of acquisition, length of residence, proficiency or offline gender knowledge. Additionally, a within-subject comparison of stimulus modalities (written vs. auditory) shows that immersed learners may show some of the effects only in the auditory modality; in non-finite verb violations, an early native-like N400 was only present for auditory stimuli. However, modality failed to influence the response to gender. Taken together, the results confirm the persistent problems of Romance learners of Dutch with online gender processing and show that they cannot be overcome by reducing task demands related to the modality of stimulus presentation.

Keywords: second language acquisition, grammatical gender agreement, event-related potentials (ERPs), P600, modality, immersion

INTRODUCTION

Second language (L2) acquisition of many aspects of syntactic structure is known to be difficult, especially when acquisition starts later in life. A major question being debated in the literature is to what extent and under what circumstances late L2 speakers can become native-like with respect to syntax processing (e.g., Clahsen and Felser, 2006; White, 2007). The evidence is mixed; in some cases this does seem to be possible, while in other cases, it is difficult or impossible. A number of factors have been suggested to play a role in this variation, but two which have received relatively little attention are the difficulty of the target grammatical system and the potential role of modality of testing (written vs. auditory presentation). The present study investigates whether event-related potential (ERP) measures of native-likeness used in this line of research might be partially dependent on stimulus modality, as this might explain some of the inconsistency in the literature.

A structure that has frequently been used to test native-like attainment in the L2, is grammatical gender, since it has been shown to pose a major challenge to L2 learners (e.g., Hawkins, 2001; White et al., 2001; Sabourin, 2003; Blom et al., 2008). Demonstrating gender processing that is comparable to that of natives therefore forms a strong test for L2 syntax acquisition. Grammatical gender is a classification system for nouns (e.g., masculine and feminine in French, or masculine, feminine and neuter in German) which allows speakers to establish syntactic cohesion between the elements in a phrase through agreement. Because the gender of a word is typically not predictable from its meaning, learning grammatical gender involves acquiring both the knowledge of a word's gender (gender assignment) and of how gender is expressed syntactically (gender agreement or concord). Therefore, L2 learners must tag each new lemma with its corresponding gender and learn which grammatical elements in the context have to agree with it. For example in Dutch, all nouns are assigned to either the common or the neuter gender class and gender concord occurs with determiners and pre-nominal adjectives (e.g., *de*$_{[def,\ common]}$ *tuin*$_{[common]}$, the garden, *een*$_{[indef]}$ *mooie*$_{[indef,\ common]}$ *tuin*$_{[common]}$, a beautiful garden). During processing, a comprehender must retrieve the noun's gender fast enough to establish gender concord. The question is (a) whether L2 learners manage to do so, and (b) whether they achieve this using the same processing strategies as native speakers.

Gender processing in L2 has already been the topic of numerous investigations using behavioral measures, such as grammaticality judgments, sentence-picture matching, (elicited) production, and eye tracking (for overviews, see, e.g., Grüter et al., 2012; Hopp, 2013). More recently, researchers have begun to employ ERPs to investigate native-likeness of grammatical

gender processing in the L2, because ERPs are known to be highly sensitive to the immediate, unconscious on-line detection, and processing of linguistic anomalies (e.g., Osterhout and Holcomb, 1992; Molinaro et al., 2011). Studies using off-line behavioral measures (e.g., White et al., 2001, 2004; Franceschina, 2005) cannot give access to this sort of evidence, which makes interpretation of their results more difficult. Some online techniques such as eye tracking (Dussias, 2010) measure real-time language processing, but do not provide us with the qualitative evidence of potential brain mechanisms that ERPs can. The rationale of such ERP studies is that the more similar the response between native speakers and learners, the more similar the underlying neural and cognitive processing mechanisms. In other words, a comparison of ERPs in native speakers and L2 learners can tell us how native-like the latter really are.

In first language processing, gender and other (morpho)syntactic violations are found to be associated with two primary kinds of components: the left anterior negativity (LAN) and the P600. The LAN has been widely associated with morphosyntactic agreement processes (Münte et al., 1993; Friederici et al., 2000; Molinaro et al., 2011), but others claim that it is a more general index of working memory load (Kluender and Kutas, 1993; Coulson et al., 1998). The P600 has been reported for a range of syntactic and other linguistic violations (e.g., Osterhout and Holcomb, 1992; Hagoort et al., 1993; Münte et al., 1993; Burkhardt, 2007). Given the extremely heterogeneous conditions that elicit a P600, this component cannot be exclusively associated with agreement specifically, or even syntactic processing difficulties more generally, and is therefore often interpreted as a late stage of (re)analysis of information (Osterhout and Holcomb, 1992; Bornkessel-Schlesewsky and Schlesewsky, 2008). It may even reflect a more general process, such as the P300 (Gunter et al., 1997; Coulson et al., 1998; but see Osterhout and Hagoort, 1999; Frisch et al., 2003). There is however, a strong correlation between the appearance of the P600 effect and grammatical violations. In contrast, findings are more varied with respect to the presence of a LAN. In addition to the LAN and P600, some studies have found an N400, or a biphasic N400-P600 pattern (but no LAN) in response to syntactic violations (see an overview reported in Molinaro et al., 2011). This is surprising, since the N400 is a component normally associated with difficulty in semantic integration (see Kutas and Federmeier, 2011, for an overview). It has therefore been proposed that an N400 in response to syntactic agreement anomalies is likely to be a result of non-syntactic information that is needed to process the mismatch, for example information that requires lexical access (Molinaro et al., 2011). Because the LAN and N400 are variable in studies of native processing, particularly for gender agreement, we will consider the P600 to be the primary measure of native-likeness, although we will report findings in the time window associated with the LAN/N400 (300–500 ms after presentation) as well.

ERP results regarding grammatical gender processing in the L2 have provided mixed results. A number of studies find that, at least under some conditions, sufficiently proficient L2 learners are able to show native-like ERP responses to gender violations. A set of studies investigating L2 processing of French suggests that English, German, and Spanish learners of French can show native-like ERP responses in the form of a P600 effect (Frenck-Mestre et al., 2009; Foucart and Frenck-Mestre, 2011, 2012). The same goes for English and Chinese learners of Spanish (Tokowicz and MacWhinney, 2005; Gillon Dowens et al., 2010, 2011). German and Polish learners of Dutch can also show a P600 in response to gender violations (Sabourin and Stowe, 2008; Loerts, 2012). Despite these consistent results, however, it is clear that this does not generalize to success in all aspects of gender processing, as the English and German learners also failed to respond in a native-like manner to gender in some forms of agreement (Foucart and Frenck-Mestre, 2011, 2012). Stronger yet, Romance learners of Dutch did not show sensitivity to gender agreement anomalies in the form of a P600 effect even in straightforward determiner noun agreement structures (Sabourin and Stowe, 2008). It is unclear why this group failed to exhibit the majority pattern; we will discuss some factors which might have affected their success in somewhat more detail.

One of the factors which has been considered to be central for native-like learning of a late L2 is whether a grammatical element (e.g., gender) is present in the L1. Many studies have focused on this question, but have reached different conclusions. There is some evidence that having a gender system in the L1 might be an advantage when acquiring an L2 gender system (e.g., Bruhn de Garavito and White, 2000; Hawkins, 2001; Franceschina, 2005). This is in favor of models proposing that the L1 restricts L2 acquisition (Hawkins and Chan, 1997). However, there is also evidence of L2 learners without gender systems in their L1 being able to show full acquisition of grammatical gender (White et al., 2001, 2004), which is seen as evidence against such a restriction (Schwartz and Sprouse, 1994, 1996; see also White, 1989; White et al., 2004). The presence vs. absence of gender in the L1 seems at the least to be more complicated than these views suggest, however.

The French and Spanish studies mentioned earlier show that learners with no gender in their L1 (English and Chinese speakers) can show native-like ERP responses. Further, Sabourin and Stowe (2008) find differences between two L1s which both have gender: German on the one hand and Romance learners on the other. Sabourin and Stowe themselves attribute their results to the (lack of) similarity between the native and target language of these learners: Dutch gender is in general predictable from the gender of the cognate German word due to their common historical origin, while there is no one-to-one-correspondence between Romance and Dutch gender at the lexical level. Moreover, agreement between noun and adjective is more similar in German and Dutch than the Romance languages and Dutch. Sabourin and Stowe conclude that processing routines are transferred from L1 to L2, rather than transfer of the abstract knowledge that nouns have gender, and that these routines must be similar for successful transfer (see Foucart and Frenck-Mestre, 2011, for a similar argument).

However, an explanation which assumes that similar routines in L1 are necessary for native-like processing does not account for the results of other studies mentioned above showing that even with no gender system in the L1, learners are able to show native-like effects. A different approach to the effects of L1

transfer is formulated within the Competition model (see Bates and MacWhinney, 1987). According to the competition-based account, when L1 does not contain gender there is no interference. This predicts successful outcomes for languages with no gender (Tokowicz and MacWhinney, 2005). However, when existing processing routines are transferred, they will cause interference if they are dissimilar from those required for L2 (accounting for the failure of the Romance learners of Dutch).

The target language itself may also contribute to the failure of Sabourin and Stowe's (2008) Romance group to show native-like processing. Most of the successful studies have investigated Romance target languages. Unlike Romance or Slavic languages, which have transparent gender systems (i.e., a predictable gender category based on morphophonological patterns), Dutch is generally regarded as having an opaque gender system (Corbett, 1991; van Berkum, 1996). Although some morphological forms predict the gender of the word, these cues are only available for a relatively small proportion of the vocabulary in the language. This clearly presents a more difficult problem for the learner than gender in a more transparent language, which may certainly explain why the Romance group in the Sabourin and Stowe study failed to achieve a native-like level.

Neither L1 interference nor target language opaqueness, however, entirely accounts for the results found by Loerts (2012). Her study demonstrates that highly advanced Polish learners of Dutch can show somewhat weak, but native-like ERP responses, even though Polish agreement differs from Dutch. Loerts' results also show that an opaque system can be learned, although it may be more difficult to learn than a transparent system. Only her most proficient learners showed native-like processing (see Davidson and Indefrey, 2009, for another example of relatively low proficient learners failing to show native-like effects for gender processing in an opaque L2 system), while even fairly low proficient English learners of Spanish have been shown to respond with a clear P600 effect (Tokowicz and MacWhinney, 2005). An alternative explanation is thus that Sabourin and Stowe's (2008) Romance learners were simply not proficient enough to show online processing comparable to that of natives. Although the proficiency of the Romance group was not investigated in detail, a similar group of German learners did significantly better when tested on offline gender knowledge (Sabourin, 2003). The Romance participants in the ERP study also performed worse at the end of sentence grammaticality judgments collected during the ERP session. It has been shown that proficiency affects brain responses (e.g., Steinhauer et al., 2006; McLaughlin et al., 2010). A replication of the Sabourin and Stowe study with a group of learners as proficient as in the Loerts study can demonstrate whether this is the sole explanatory factor. This is one of the aims of the current study.

However, there is another factor that may have produced the difference between the two Dutch studies, which has thus far been overlooked: testing modality. Unlike virtually all the other studies summarized above, Loerts (2012) tested her Polish learners using auditory sentence presentation. She argues that the learners had acquired their L2 primarily in the auditory modality as emigrants who arrived with no formal training in their new language. Consequently, processing routines may be tuned to the auditory stimulus modality. Indeed, the experience of learning in immersion can be expected to differ substantially from a formal learning environment. Yet, the various populations that have been tested so far differ in this domain. The participants in the Romance studies summarized above included learners with extensive formal training in their L2. In many of the studies there was no immersion (Tokowicz and MacWhinney, 2005; Gillon Dowens et al., 2011) or only minimal immersion during the participants' recent residence in France (Foucart and Frenck-Mestre, 2011, 2012). Sabourin and Stowe (2008), unlike Loerts, tested a similar late immersion population using visual materials, with each word presented consecutively in the center of the screen. An alternative explanation for the lack of a native-like response in their study could thus be difficulties with the visual presentation. Below, we will speculate about why a visual ERP paradigm might, under some circumstances, be problematic.

In a typical language comprehension ERP paradigm, participants are presented with sentences displayed one word at a time at the center of a screen, at a rate of around two words per second, a technique called rapid serial visual presentation (RSVP). The advantages of this method are that the duration of stimulus presentation can be controlled (and manipulated) tightly, that eye movements, which lead to large artifacts in the EEG, are reduced to a minimum, and that making the stimulus material and time-locking the brain responses to the presentation of violations in the stimulus is relatively straightforward. Consequently, a large majority of ERP sentence comprehension studies use this method. In contrast, auditory sentence presentation is used much less frequently in ERP research. With spoken stimuli, it is more difficult to control the presentation duration of individual words. In addition, making recordings of spoken sentences is more time consuming and requires tight control of acoustic confounds (e.g., prosodic cues about upcoming information, Dimitrova et al., 2012), as well as timing issues (e.g., setting markers to millisecond precision for the events of interest).

We do not expect to find interesting differences between word-by-word reading and listening for language processing in natives (Müller et al., 1997; Hagoort and Brown, 2000; Balconi and Pozzoli, 2005). In the L1, learners develop fully automatized processing of both modalities; moreover, the auditory representation of language is automatically activated by written materials (Perfetti et al., 1992; Frost, 1998), so that the routines activated during auditory processing can be utilized as well as those specific to the written modality (Homae et al., 2002). Despite expecting comparable results for the two modalities in general, even for L1 comprehenders, consecutive word by word presentation in the middle of the screen presents a challenge under some circumstances. The optimum speed of presentation is an issue; Hopp (2010) shows that speeded RVSP presentation can make even native speakers break down in their grammaticality judgment ability, making their performance mirror that of L2 learners (see also Camblin et al., 2007, who show a case where speeded RSVP eliminates an effect which is clear in naturally produced connected speech). Conversely, studies directed at optimizing computerized text presentation on small screens have shown that too slow a presentation can also interfere with comprehension (Bernard et al., 2001). This may result from working memory and

maintenance issues. Stowe (1991) showed that readers were more likely to garden-path or have difficulty in recovering from a garden path with center of the screen presentation, as opposed to presentation of words across the screen in their normal position, even when readers were allowed to pick their optimum pace.

L2 learners differ in a number of ways from native speakers, some of which can be expected to interact with modality. First, their cumulative reading experience in the L2 is likely to be substantially lower than that of native speakers. This means that their activation of the L2 via this modality can be expected to be less automatized than in native speakers (Koda, 1996). Second, interference from the writing system of the first language may lead to even less activation of the phonological form of the L2, in comparison with natives (Koda, 1999). These differences can potentially play a role for all L2 learners, but may be especially relevant for learners with less formal instruction in the language and in whom learning took place primarily via the auditory modality. The optimum speed of presentation is also likely to differ between various groups of learners and natives. This issue has received relatively little attention in the literature, but given that stimulus modality was one difference between the unsuccessful Romance group reported by Sabourin and Stowe (2008) and the relatively more successful group studied by Loerts (2012), this factor was included in the current experiment in order to determine whether it explains the different patterns seen in the two studies. A clear effect of modality would suggest that researchers need to pay more attention to this variable in their experimental designs, and might have implications for the differences between immersed and instructed learners as well.

Summarizing, the goal of the current study is to gain more insight into why some groups may show persistent problems in attaining native-like processing of grammatical gender. We investigate grammatical processing in immersed Romance L2 learners of Dutch, using the P600 as a measure of native-likeness, in order to answer the question whether late L2 learners can show native-like syntactic processing, even if the gender marking in the L1 differs from that in the L2, which may cause interference, and the L2 gender system is relatively opaque, making it harder to recognize the grammatical agreement regularities. Following Sabourin and Stowe (2008), in addition to gender violations, which have proven difficult to master, we present our participants with non-finite verb violations, a construction that is relatively easy to acquire, as a baseline for comparison. We compare the responses of high-proficient Romance learners with those of native speakers of Dutch. Additional measures of proficiency will be gathered from the first. A within-subject comparison of stimulus modalities allows us to determine whether the absence of a P600 effect for gender in the Sabourin and Stowe (2008) study was due to processing demands associated with the task modality.

In addition to standard group analyses of the ERP waveforms, we will closely inspect individual differences within each group. Adding these analyses has several benefits. First, lack of effects in grand mean ERP results does not necessarily mean that none of the individuals showed a native-like ERP response. Rather, a null effect might be based on opposite effects (a positive going effect in one set of individuals and a negative going effect in others) canceling each other out. In a similar way, biphasic responses can be a spurious result of averaging (Osterhout, 1997; Nieuwland and Van Berkum, 2008; Tanner and Van Hell, 2014; Tanner et al., 2014). Before we draw any strong conclusion that a group of learners' processing of gender agreement qualitatively differs from natives, it is important to identify varying patterns in each of the groups. Furthermore, there may be predictors of native-likeness in L2 learners, such as age of acquisition, proficiency, language exposure and use, that may explain variance within the group (e.g., Weber-Fox and Neville, 1996, 1999; Rossi et al., 2006; Steinhauer et al., 2009; Tanner et al., 2014). Understanding which individual difference factors, if any, are associated with the outcome in L2 learning is a fundamental question which is difficult to answer with group-based analyses, and might also help us determine the source of some of the mixed patterns of results in L2 gender research.

MATERIALS AND METHODS

PARTICIPANTS

Participant characteristics and proficiency scores can be found in **Table 1**. Forty-five participants took part in the experiment. Seven participants had to be excluded from the analyses because of too many artifacts in the EEG signal. Nineteen of the remaining participants were Romance learners of Dutch (six French, five Italians, three Romanians, five Spanish). The remaining 19 participants were native speakers of Dutch. All participants were right handed, neurologically unimpaired and did not have any problems with hearing, speaking, or writing. Prior to conducting any procedures, written consent was obtained from all participants for the study, which was approved by the local ethics committee. Participants were fully debriefed at the end of the experiment and received a small fee for participation.

All learners had moved to the Netherlands at or after the age of 16 and had been immersed in the L2 context for at least 5 years at the time of testing. The learners had very little to no exposure to Dutch before immigration. They were asked to indicate the frequency of use of Dutch in daily life: a composite score for L2 use was calculated based on questions about language use at home (with partner and children), outside of the home (at the workplace and other), and use of Dutch media. They additionally answered questions about their use of Dutch in a specific modality: they estimated the percentage of use of the L2 in the visual modality (i.e., reading/writing) compared to the auditory modality (i.e., speaking/listening), both during learning of Dutch at onset of immigration and during everyday life at the time of testing.

L2 proficiency was assessed by means of several (written) measures. A pre-selection on the basis of a pre-test in the form of 20 grammar items of the Dutch DIALANG Placement Test (adapted from http://www.lancaster.ac.uk/researchenterprise/dialang/about.html) ensured that all participants had a relatively high level of proficiency in Dutch. Participants had to complete at least 13 of the items correctly to be selected for participation. Another proficiency measure was taken in the lab, in the form of a C-test (constructed by Keijzer, 2007), which consisted of two texts containing gaps where parts of some words had been left out. The participants' task was to fill the gaps. After the EEG experiment, participants were also asked to complete

Table 1 | Means (and ranges) of participant characteristics and scores on proficiency measures, and significance of between-group comparisons (Mann-Whitney *U*-test).

Measure	Learners ($n = 19$)	Natives ($n = 19$)	*U*- and *p*-value
AGE/EXPOSURE/USE			
Age at testing (years)	42.3 (24–64)	39.8 (21–59)	$U = 162$, $p = 0.599$
Age of acquisition (years)	26.0 (16–39)	–	–
Length of residence (years)	16.3 (5–43)	–	–
L2 use (%)[a]	58.4 (12.3–87.3)	–	–
USE OF MODALITY: DURING LEARNING (%)[b]			
Visual	43.7 (20–70)	–	–
Auditory	56.3 (30–80)	–	–
USE OF MODALITY: CURRENT (%)[c]			
Visual	42.6 (20–70)	–	–
Auditory	57.4 (30–80)	–	–
PROFICIENCY MEASURES			
C-test (%)[d]	79.4 (42.1–100)	95.2 (68.4–100)	$U = 299.5$, $p < 0.001$
Gender assignment task (%)[e]	87.3 (64.6–100)	99.5 (93.8–100)	$U = 332.5$, $p < 0.001$
SELF-RATED PROFICIENCY[f]			
Reading	4.4 (3–5)	–	–
Writing	3.6 (1–5)	–	–
Speaking	3.9 (2–5)	–	–
Listening	4.3 (3–5)	–	–

[a] *Composite score based on language use inside and outside of the home and use of Dutch media.*

[b] *Percentage of L2 use in the visual modality (i.e., reading/writing) compared to the auditory modality (i.e., speaking/listening) during learning of Dutch at onset of immigration.*

[c] *Percentage of L2 use in the visual modality (i.e., reading/writing) compared to the auditory modality (i.e., speaking/listening) in everyday life at the time of testing.*

[d] *Percentage of correct responses on the C-test (spelling errors were not penalized).*

[e] *Percentage of correct responses (i.e., a minimum of 2/3 instances of each item assigned correctly) on the gender assignment task.*

[f] *Ratings on a 5-point scale with five as highest level of skill in Dutch.*

an offline gender assignment task. This task was used to test the participants' knowledge of the grammatical gender of the critical nouns used in the EEG experiment. In addition to these measures, learners rated their L2 Dutch in terms of reading, writing, speaking, and listening proficiency on a Likert-scale between 1 (very bad) and 5 (very good). Participants' scores on the proficiency measures can be found in **Table 1**.

MATERIALS

The design and materials of the EEG experiment were largely based on work by Loerts (2012), who studied L2 gender and non-finite verb processing in natives and Slavic learners of Dutch. One hundred and forty-four experimental sentences were created (see **Table 2** for examples, the full list of sentences can be found in the Supplementary Material, Data Sheet 1). Forty-eight of the sentences[1] were used to test non-finite verb agreement. Half of these contained an infinitive and the other half a past participle verb. For their ungrammatical counterparts, these verbs were altered into their participial or infinitival form, respectively. The other 96 sentences were used to test grammatical gender agreement. In these sentences, the determiner either agreed in gender with the following noun or violated gender concord. Determiner and noun were either adjacent, or non-adjacent (with an adjective intervening between the determiner and noun). Only highly frequent Dutch target nouns and verbs were used (nouns: mean = 2.16, range = 0.78–3.08; verbs: mean = 2.46, range = 0.95–4.05, on log lemma frequency of occurrence per million taken from the CELEX corpus: Baayen et al., 1995). Finally, 122 well-formed filler sentences were included. These filler sentences were added to raise the overall proportion of correct sentences to about 3/4, making the task more similar to natural language processing.

For the auditory part of the experiment, spoken forms of all sentences were recorded. Each sentence was read aloud by a female native speaker with a standard Dutch accent who was trained to produce correct and incorrect sentences with normal intonation. Despite training, acoustic confounds, such as subtle prosodic cues to the upcoming ungrammaticality remain possible (Dimitrova et al., 2012). To prevent any influence of such confounds, each sentence was presented in its original form or in a digitally spliced version, constructed by cross-splicing the original recordings of grammatical and violation sentences, cutting at the onset of the determiner for the gender condition, or the verb

[1] Because of the large number of factors in the current design, it was not possible to get a high number of trials per condition without making the experiment too long, which in all probability would have resulted in severe fatigue effects in our data. We realize that as a result, the number of trials per condition is on the low side, particularly for the non-finite verb condition. However, highly salient agreement errors, such as the non-finite verb agreement violations used in the current study, have been shown to elicit large ERP effects. As the results section of this paper shows, even with this low number of trials we had sufficient power to find significant effects in this condition. In the less salient gender condition however, there was double the amount of trials per condition to ensure sufficient power.

Table 2 | Example materials of the EEG experiment.

Condition	Example sentences	Number of items per list
Non-finite verb agreement	Ze heeft alleen haar beste vriendin uitgenodigd/*uitnodigen voor haar verjaardag. (She has only invited/*invite her best friend for her birthday.) Hij probeert me altijd aan het lachen te maken/*gemaakt door grapjes te vertellen. (He always tries to make/*made me laugh by telling yokes.)	12/12 visual, 12/12 auditory
Gender agreement	Vera plant rode rozen in de/*het tuin van haar ouders. (Vera is planting red roses in the$_{com}$/*the$_{neu}$ garden of her parents.) Het duurde uren voordat Jeroen het/*de nette pak van zijn broer had aangetrokken. (It took hours for Jeroen to put on the$_{neu}$/*the$_{com}$ fancy suit of his brother.)	24/24 visual, 24/24 auditory

Critical targets, where the ERP was measured, are underlined.

in the non-finite verb condition. Noise reduction and volume normalization were applied to all sound files.

A within-subject design was employed to test the effects of modality within the same group of subjects. Eight experimental lists were created using a Latin Square design, crossing the factors modality (visual, auditory), correctness (correct, incorrect), and splicing (spliced, unspliced), to ensure each participant was presented with only one version of each sentence and an equal number of each type. Each list was presented to two or three participants from each group, and each participant saw only one list.

PROCEDURE

Event-related potentials were recorded while participants listened to or read the sentences. After each sentence, the participant had to make a grammaticality judgment. Participants were comfortably seated in an electrically shielded and sound attenuated chamber. The sentences were presented using E-prime (Schneider et al., 2002a,b), which in addition recorded accuracy with respect to the grammaticality judgments. Visual stimuli were presented on a computer screen in front of the participants. Speakers were placed to the left and right side of the screen. Visual sentences were presented at a rate of two words per second: each word was presented for 250 ms, followed by 250 ms blank screen. Auditory sentences were presented at normal speech rate. Participants were asked to avoid moving any parts of their body and not to move their eyes or blink during sentence presentation. The experiment consisted of four blocks: either two visual blocks followed by two auditory blocks or the reverse. The duration of the breaks between blocks was determined by the participant. Altogether, the EEG experiment lasted about 1 h.

Subsequently, participants were asked to fill in the pen and paper C-test. Finally, they performed a gender assignment task on a computer. The target words of the EEG experiment were presented in randomized order, each item appearing three times. Participants were instructed to indicate, by a mouse click on either the common ("de") or neuter ("het") definite article, whether they thought the word had common or neuter gender in Dutch.

EEG RECORDING AND ANALYSIS

The continuous EEG (500 Hz/22 bit sampling rate) was recorded from 54 Ag/AgCl scalp electrodes mounted into an elastic cap (Electro Cap International, Inc.) according to the international extended 10–20 system (see **Figure 1** for recording sites). To monitor eye-movements, four additional electrodes were placed on the outer canthi of each eye and above and below the left eye. Scalp electrode signals were measured against a common reference during recording. Impedances were reduced to below 10 kΩ[2]. The amplifier (TMS international) measured DC with a digital FIR filter (cutoff frequency 130 Hz) to avoid aliasing. After acquisition, the raw data were further processed with Brain Vision Analyzer 2.0.4. The data were re-referenced to the average of two electrodes placed over the left and right mastoids and digitally filtered with a high-pass filter at 0.1 Hz and low-pass filter at 40 Hz. The data were segmented, time-locked to the onset of the critical target (from 500 ms before to 1400 ms

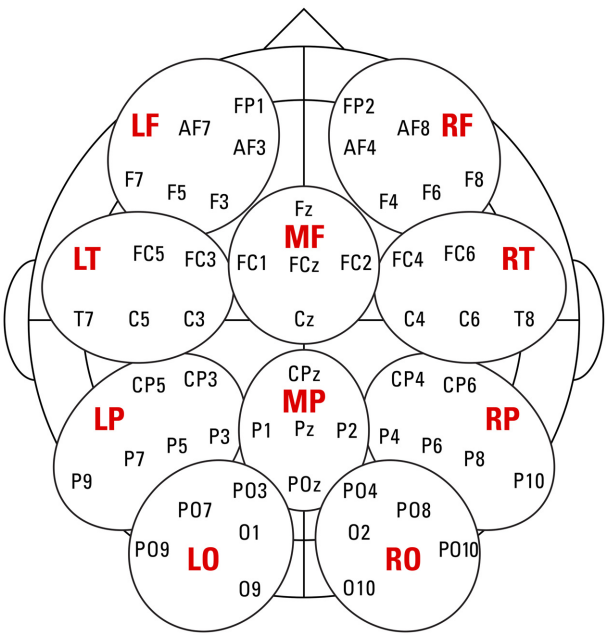


FIGURE 1 | Approximate location of the recording sites and the 10 regions of interest used for analyses: left/middle/right frontal (LF/MF/RF), left/right temporal (LT/RT), left/middle/right parietal (LP/MP/RP), and left/right occipital (LO/RO).

[2]In some instances, some temporal and frontal electrodes could only be reduced to below 20 kΩ.

after stimulus onset). Average ERPs were formed without regard to behavioral responses, from trials free of muscular and ocular artifacts; the latter were corrected using the Gratton and Coles procedure (1989). Individual channel artifacts led to rejection of 0.5% of the data in the learner group and 0.6% in the native group. A baseline period was set from 200 to 0 ms before onset of the critical words to normalize the data. A total of 10 regions of interest (ROIs), containing five or six electrodes each, were used for analyses (depicted in **Figure 1**).

We analyzed amplitudes of the ERP waveforms in the time-windows in which a LAN/N400 and P600 are to be expected: 300–500 and 600–1200 ms after stimulus onset. The latter window is somewhat longer than is typical in P600 studies in monolinguals, because the P600 in L2 learners can be somewhat delayed (Weber-Fox and Neville, 1996; Hahne, 2001; Rossi et al., 2006; Sabourin and Stowe, 2008). For grand mean analyses, ANOVAs were calculated within each time window and *sentence structure* (non-finite verb, grammatical gender) separately, using the ezANOVA function of the ez package (version 4.2.2: Lawrence, 2013), implemented in R (version 3.1.0: R Core Team, 2014). The analyses included *correctness* (grammatical, violation) and *modality* (visual, auditory) as within-participants factors, and *group* (natives, learners) as between-participants factor. Data from lateral (left and right frontal, temporal, parietal, and occipital ROIs) and medial (middle frontal and middle parietal ROIs) regions were treated separately in order to identify topographic and hemispheric differences. For the lateral regions, the ANOVA also included *hemisphere* (left, right) and *anterior-posterior* (frontal, temporal, parietal, occipital) as within participants factors. For the medial regions, *anterior-posterior* (frontal, parietal) was the only topographical factor in the ANOVA. The Greenhouse-Geisser correction was applied for violations of the sphericity assumption. Only main effects of, and interactions with, *correctness* are reported. In the presence of a significant higher-level interaction, lower-level interactions, and main effects are not interpreted. False discovery rate correction (Benjamini and Hochberg, 1995) was applied for follow-up tests to control for Type 1 error. Additional regression analyses, performed in R version 3.1.0 using the lm function of the lme4 package (version 1.1.6: Bates et al., 2014) will be described together with the results.

RESULTS

BEHAVIORAL RESULTS

The percentages of accurate grammaticality judgments per *group*, *sentence structure*, and *modality* are shown in **Figure 2**. A Three-Way ANOVA was conducted on the arcsine transformed proportions of correct responses to stabilizes variance and normalize the data (mean and SDs reported below are from the untransformed percentages). The ANOVA revealed a significant main effect of *group*, $F_{(1,\ 36)} = 53.24$, $p < 0.001$, with the learners giving fewer correct responses than the natives (mean = 71.1, $SD = 17.8$ vs. mean = 93.0, $SD = 11.5$). The main effect of *sentence structure*, $F_{(1,\ 36)} = 41.66$, $p < 0.001$, shows that the average performance is worse in the gender condition. However, there is also a significant interaction between *group* and *structure*, $F_{(1,\ 36)} = 5.55$, $p = 0.024$. Paired comparisons show that the difference between structures is highly significant in the learner group [$t_{(62.9)} = 4.91$, $p < 0.001$, gender mean = 62.8, $SD = 14.1$; non-finite verb mean = 79.5, $SD = 17.3$]. There is a smaller, but still significant difference between structures in the native group [$t_{(59.7)} = 2.42$, $p = 0.019$, gender mean = 92.2, $SD = 6.2$; verbs mean = 93.8, $SD = 15.1$]. Interestingly, with respect to one of our research questions, there is a significant main effect of *modality*, $F_{(1,\ 36)} = 8.37$, $p = 0.006$, with the percentage of correct responses in the auditory condition being somewhat lower than in the visual condition (mean = 79.5, $SD = 20.0$ vs. mean = 84.6, $SD = 16.7$). There are however no significant interactions between *modality* and *group*, *modality* and *structure*, or *group*, *modality,* and *structure* (all Fs < 3).

ERP RESULTS: GRAND MEAN ANALYSES

Figures 3, **4** show the grand mean ERP waveforms for natives and learners, respectively. Results of the omnibus ANOVAs are provided in the Supplementary Material (Data sheet 2). Significant results and follow-up analyses will be described below.

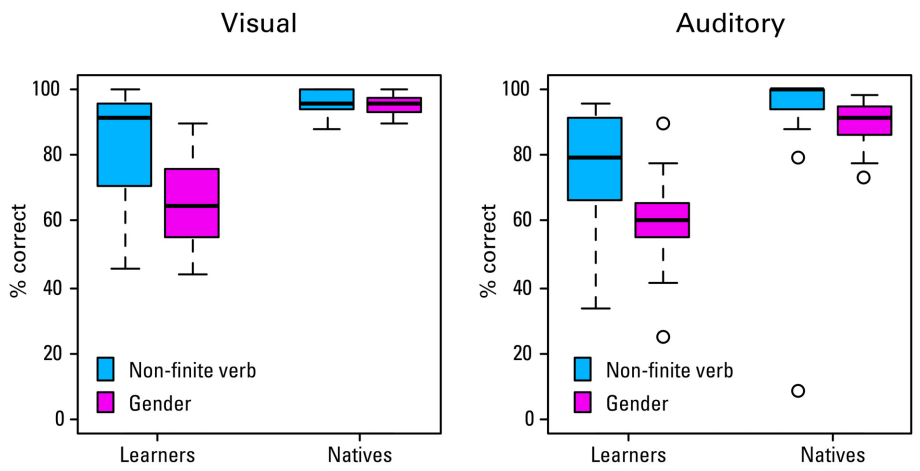


FIGURE 2 | Accuracy on grammaticality judgments made during ERP recording session by group, modality, and structure.

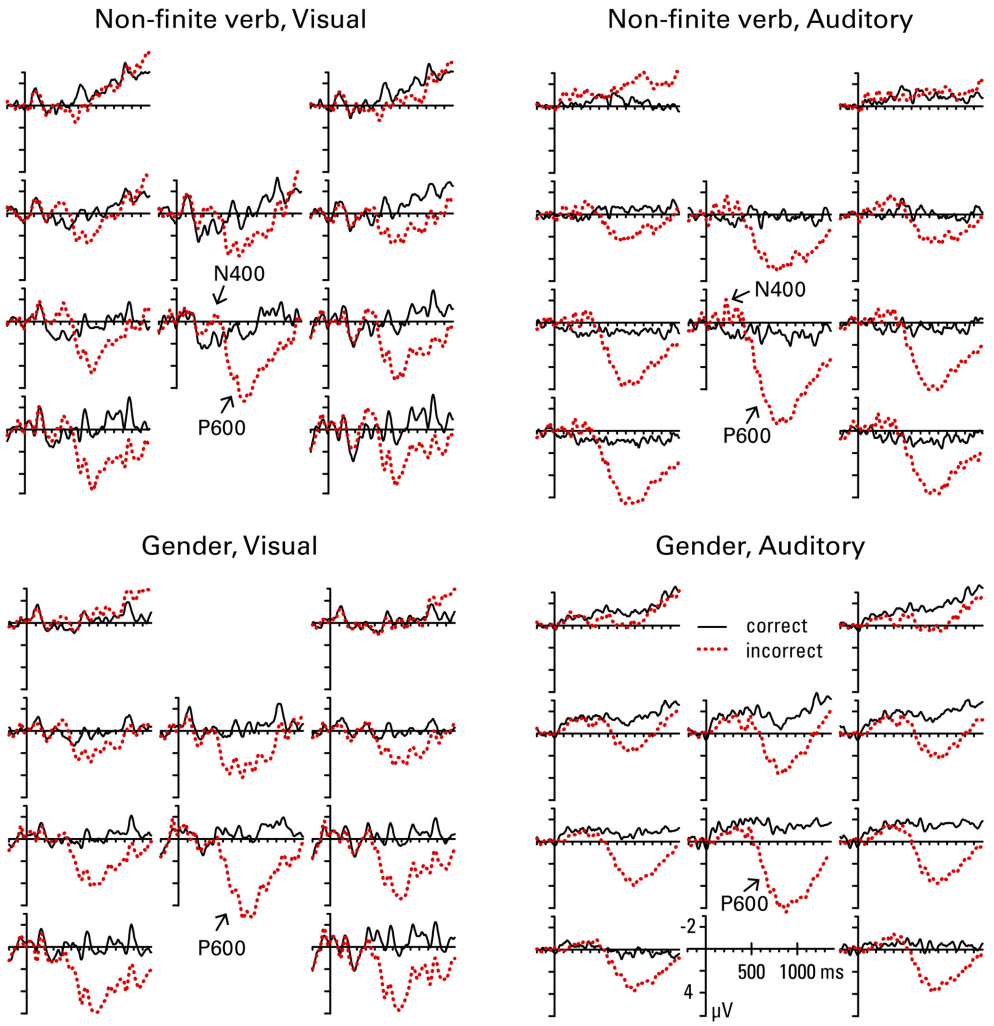


FIGURE 3 | Natives' grand average ERP waveforms at all 10 regions of interest (see Figure 1) for correct and incorrect use of non-finite verb and gender agreement in the visual and the auditory condition.

Non-finite verb agreement

In the 300–500 ms window, the lateral omnibus ANOVA for the non-finite verb condition showed a significant *correctness* by *anterior-posterior* interaction, $F_{(3,\ 108)} = 6.02$, $p = 0.011$; follow-up analysis revealed that the effect of *correctness* reached significance in posterior regions only [frontal, $F_{(1,\ 36)} = 0.52$, $p = 0.476$; temporal, $F_{(1,\ 36)} = 4.16$, $p = 0.065$; parietal, $F_{(1,\ 36)} = 14.70$, $p = 0.002$; $F_{(1,\ 36)} = 11.77$, $p = 0.004$], with the incorrect condition showing more negative voltages than the correct condition. Due to a marginally significant *group* by *correctness* interaction in the omnibus ANOVA, $F_{(1,\ 36)} = 3.65$, $p = 0.064$, another follow-up analysis was conducted separately for natives and learners. This analysis revealed that the main effect of *correctness* was significant in natives, $F_{(1,\ 18)} = 7.36$, $p = 0.028$, but not in learners, $F_{(1,\ 18)} = 0.08$, $p = 0.780$. The medial omnibus ANOVA revealed a significant main effect of *correctness*, $F_{(1,\ 36)} = 9.22$, $p = 0.004$, with more negative voltages for the incorrect than the correct condition. Due to a marginally significant *correctness* by *anterior-posterior* interaction, $F_{(1,\ 36)} = 3.33$, $p = 0.076$, a follow-up analysis was conducted, which again revealed that the effect of correctness reached significance in the posterior region only [frontal, $F_{(1,\ 36)} = 2.77$, $p = 0.105$; parietal, $F_{(1,\ 36)} = 14.55$, $p = 0.002$]. Additionally, the omnibus ANOVA showed a marginally significant *group* by *correctness* by *modality* interaction, $F_{(1,\ 36)} = 3.56$, $p = 0.067$; but follow-up analyses failed to reveal a significant modality effect in either of the groups [*correctness* by *modality* interaction: natives, $F_{(1,\ 18)} = 0.72$, $p = 0.407$; learners, $F_{(1,\ 18)} = 4.12$, $p = 0.114$]. The main effect of *correctness* reached significance on its own in natives, $F_{(1,\ 18)} = 6.26$, $p = 0.044$, but not in learners, $F_{(1,\ 18)} = 3.00$, $p = 0.100$. Since visual inspection of the grand mean waveforms seems to suggest a possible negativity in medial regions for learners in the auditory condition, and finding a native-like effect in this time window for L2 learners is unusual, we performed an additional follow-up analysis separately for each modality in learners, which showed a significant *correctness* effect in the auditory, $F_{(1,\ 18)} =$

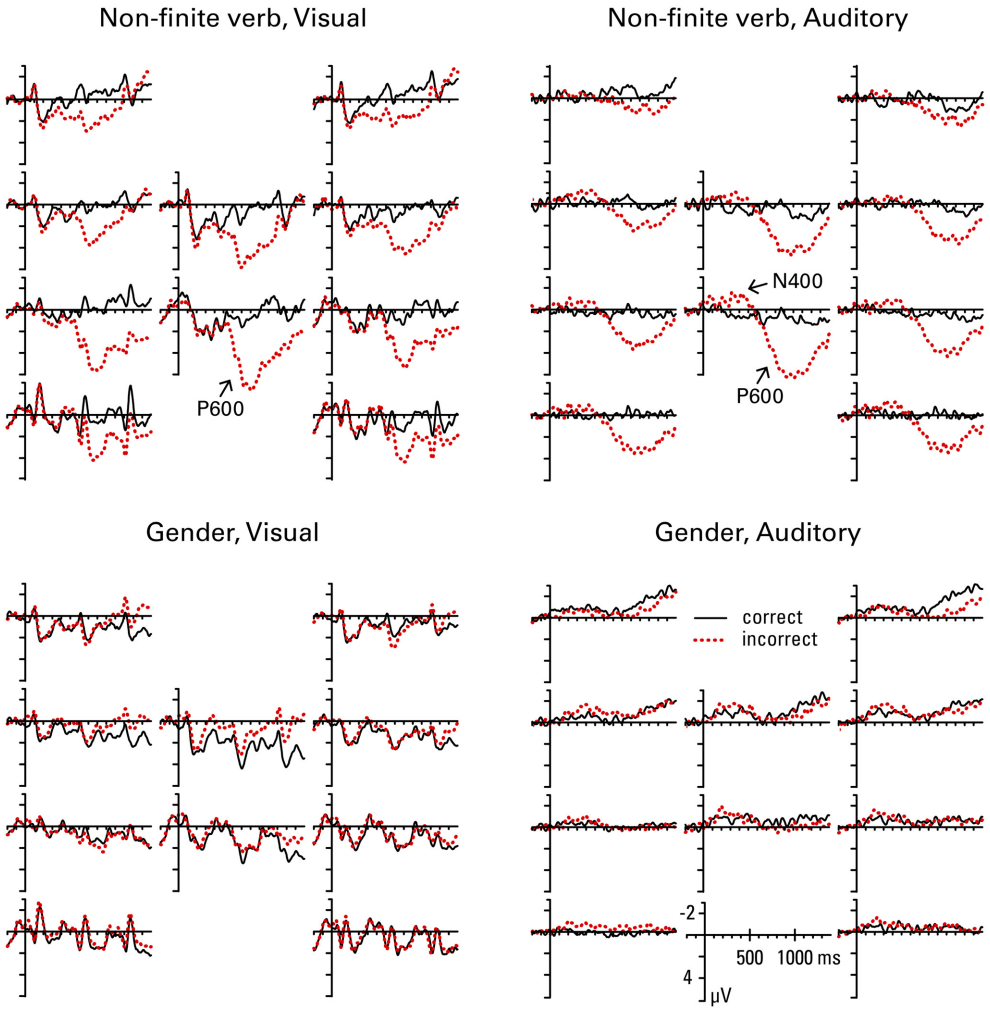


FIGURE 4 | Learners' grand average ERP waveforms at all 10 regions of interest (see Figure 1) for correct and incorrect use of non-finite verb and gender agreement in the visual and the auditory condition.

6.18, $p = 0.046$, but not the visual modality, $F_{(1,\ 18)} = 0.43$, $p = 0.522$.

In the later time window (600–1200 ms), the lateral omnibus ANOVA showed a significant *group* by *correctness* by *anterior-posterior* interaction, $F_{(3,\ 108)} = 5.95$, $p = 0.008$. Follow-up analysis revealed a significant main effect of *correctness* in both groups [natives, $F_{(1,\ 18)} = 20.39$, $p = 0.001$; learners, $F_{(1,\ 18)} = 14.16$, $p = 0.001$], with more positive amplitudes in the incorrect compared to the correct condition. A significant *correctness* by *anterior-posterior* interaction was present for natives only [natives, $F_{(3,\ 54)} = 23.51$, $p = 0.001$; learners, $F_{(3,\ 54)} = 1.97$, $p = 0.169$], which was driven by the fact that the positivity in natives was significant in the temporal, $F_{(1,\ 18)} = 16.32$, $p = 0.001$, parietal, $F_{(1,\ 18)} = 36.07$, $p = 0.001$, and occipital region, $F_{(1,\ 18)} = 35.54$, $p = 0.001$, but not the frontal region, $F_{(1,\ 18)} = 0.00$, $p = 0.985$. The medial omnibus ANOVA revealed a significant *correctness* by *anterior-posterior* interaction, $F_{(1,\ 36)} = 22.93$, $p < 0.001$; a follow-up analysis showed that the *correctness* effect is stronger in the parietal, $F_{(1,\ 36)} = 68.36$, $p < 0.001$, than the frontal region, $F_{(1,\ 36)} = 29.15$, $p < 0.001$.

It is apparent from these grand mean analyses that non-finite verb agreement violations are associated with a biphasic pattern of an N400 followed by a P600 in natives. The lack of significant effects for the frontal regions rules out a LAN effect in the 300–500 ms time window. Learners' responses are very similar to natives' in the later time-window (P600). However, in the early time window learners fail to show a native-like effect (N400) in the visual condition, and only show a smaller and less broadly distributed N400 compared to natives in the auditory condition.

Gender agreement

In the 300–500 ms window, the lateral omnibus ANOVA for the gender condition showed a significant *correctness* by *modality* by *anterior-posterior* interaction, $F_{(3,\ 108)} = 3.90$, $p = 0.039$, and a *group* by *correctness* by *modality* by *hemisphere* interaction, $F_{(1,\ 36)} = 5.24$, $p = 0.028$. Follow-up analyses conducted separately for natives and learners revealed a significant *correctness*

by *modality* by *anterior-posterior* interaction in natives, $F_{(3, 54)} = 6.28$, $p = 0.016$, but no significant effects in learners (all Fs < 2.03). However, in natives, neither the main effect of *correctness* nor the *correctness* by *anterior-posterior* interaction reached significance in either of the modalities analyzed separately (all Fs < 3.90). The medial omnibus ANOVA showed a significant *group* by *correctness* interaction, $F_{(1, 36)} = 4.30$, $p = 0.045$. However, follow-up analyses failed to find a significant main effect of *correctness*, or any of its interactions, in either of the groups analyzed separately (all Fs < 4.23).

In the 600–1200 ms window, the lateral omnibus ANOVA revealed a significant *group* by *correctness* by *anterior-posterior* interaction, $F_{(3, 108)} = 20.17$, $p < 0.001$, and a significant *correctness* by *modality* by *anterior-posterior* interaction, $F_{(3, 108)} = 7.31$, $p = 0.002$. Follow-up analyses conducted separately for natives and learners revealed a significant *correctness* by *modality* by *anterior-posterior* interaction in natives, $F_{(3, 54)} = 6.17$, $p = 0.014$, but no significant effects in learners (all Fs < 1.81). In natives, the main effect of *correctness* was significant in all regions except for the frontal one [frontal, $F_{(1, 18)} = 0.06$, $p = 0.806$; temporal, $F_{(1, 18)} = 14.33$, $p = 0.001$; parietal, $F_{(1, 18)} = 38.20$, $p = 0.001$; occipital, $F_{(1, 18)} = 35.39$, $p = 0.001$], with amplitudes in the incorrect condition being more positive compared to the correct condition. The *correctness* by *modality* interaction did not reach significance in any of the regions (all Fs < 4.03). The medial omnibus ANOVA showed a significant *group* by *correctness* by *anterior-posterior* interaction, $F_{(1, 36)} = 11.24$, $p = 0.002$. Follow-up analyses revealed that this was due to a significant *correctness* by *anterior-posterior* interaction in natives, $F_{(1, 18)} = 26.82$, $p = 0.001$, but not learners, $F_{(1, 18)} = 1.86$, $p = 0.190$. The interaction in natives was driven by the fact that the effect of *correctness* was stronger in the posterior region [frontal, $F_{(1, 18)} = 13.04$, $p = 0.002$; parietal, $F_{(1, 18)} = 47.69$, $p < 0.001$].

These grand mean analyses show that while natives show a classic P600 effect in response to gender agreement violations, learners do not: the P600 is absent for learners, in both modalities. In the early time window, there are again no effects for learners, while the natives seemed to show some small effects, which however failed to reach significance in follow-up analyses.

Figure 5 summarizes the P600 and N400 effects, showing the difference in amplitude between the violation condition and the grammatical condition, collapsed over middle frontal and all temporal, parietal and occipital ROIs, per group, structure, and modality. We see P600 effects for natives, preceded by an N400 effect in non-finite verb violations, but not gender violations. In contrast, the learners only show P600 effects for non-finite verb violations, but they do not show any effects of gender violation. The learners also show a small N400 effect for auditory non-finite verb violations (an effect that only reached significance in the medial regions).

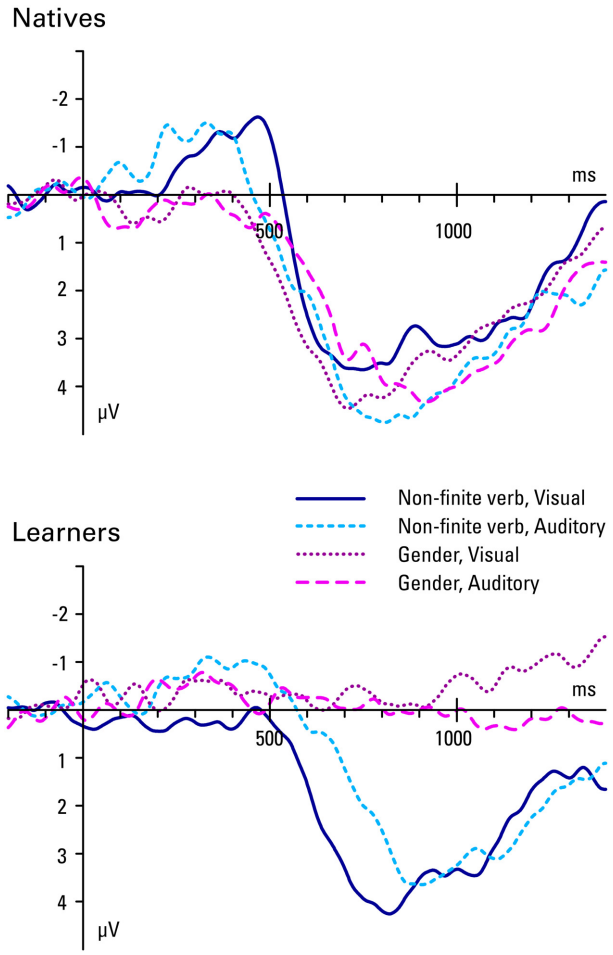


FIGURE 5 | ERP difference waves (incorrect minus correct sentence) per group, structure, and modality, collapsed over middle frontal and all temporal, parietal, and occipital ROIs.

ERP RESULTS: INDIVIDUAL DIFFERENCES ANALYSES

In this section, we will have a closer look at individual differences. First, we will investigate the distribution of N400 and P600 effects across individuals, which can be of importance for the interpretation of the grand mean results, as discussed in the Introduction. Second, we will explore possible predictors of native-likeness in the learner group, since previous research has revealed that age of acquisition, length of residence, L2 proficiency and use can affect ERP responses (also discussed in the Introduction).

Closer inspection of the N400 and P600 patterns

Following work by Osterhout and colleagues (McLaughlin et al., 2010; Tanner et al., 2013, 2014) we regressed individuals' N400 effect magnitude onto their P600 effect magnitude, to investigate the distribution of these two components across individuals. The effect magnitude here refers to the average voltage difference between conditions: correct minus incorrect in the 300–500 ms window for the N400, and incorrect minus correct in the 600–1200 ms time window for the P600. Amplitudes were averaged across middle frontal and all temporal, parietal, and occipital regions, where the N400 and P600 effects are to be expected.

Figure 6 shows the scatterplots of the results, for each group and sentence structure separately. We also investigated each modality separately, but since the results looked highly similar between modalities, these will not be discussed here. The figure informs us about whether the grand mean waveforms are

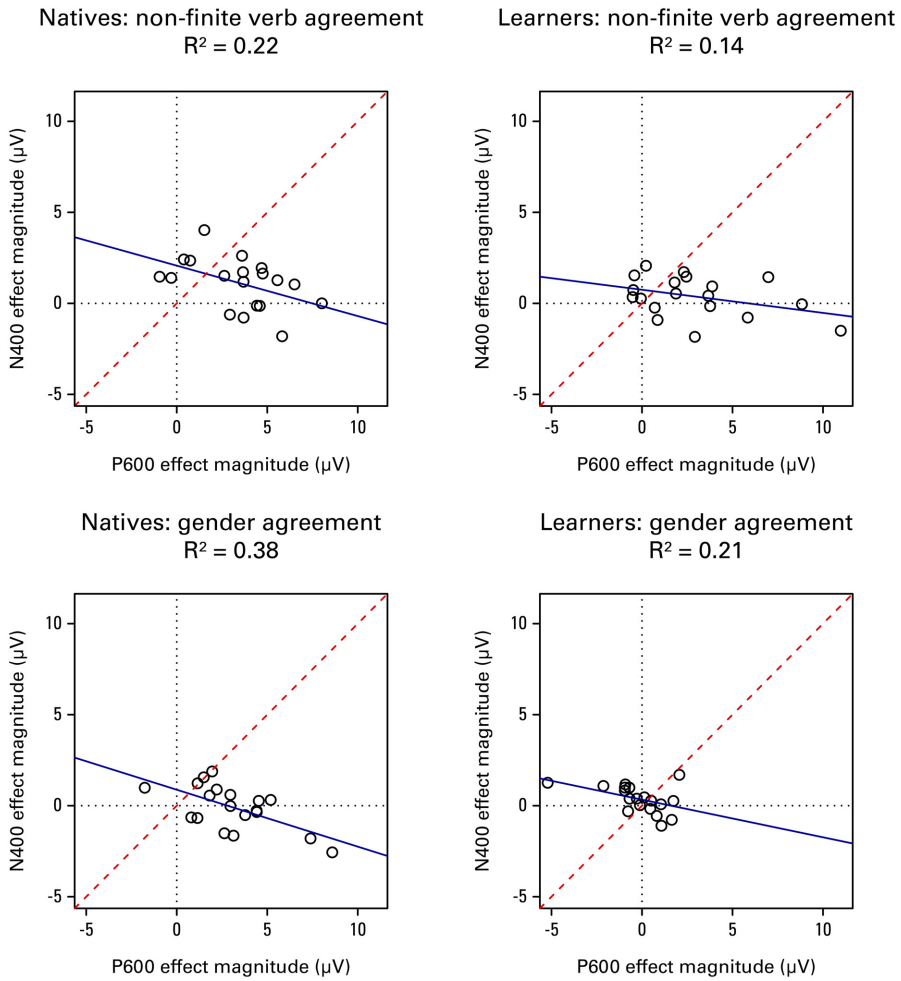


FIGURE 6 | The distribution of N400 and P600 effect magnitudes (correct minus incorrect for N400, incorrect minus correct for P600) across learners, averaged within middle frontal and all temporal, parietal, and occipital ROIs. Each dot represents a data point from a single participant. The solid line shows the best-fit regression line. The dashed line represents equal N400 and P600 effect magnitudes: individuals above/to the left of the dashed line showed primarily an N400 effect, whereas individuals below/to the right of the dashed line showed primarily a P600 effect. In the non-finite verbs many individuals show biphasic responses (upper right quadrants), whereas in the gender condition there are more sustained positivities (lower right quadrants). Very few individuals show sustained negativities (upper left quadrants). Basically none of the learners are able to show sensitivity to gender violations.

representative of most individuals' ERP profiles. We concluded from our grand mean analyses that natives show a biphasic N400-P600 pattern for non-finite verb violations, and only a P600 for gender agreement violations. Examining **Figure 6** we indeed see that the biphasic pattern is present for the majority of individuals in the non-finite verbs, and that a P600 (without preceding N400) is dominant for gender. The grand mean results of the learners showed native-like effects for verbs, but not for gender. This conclusion still holds if we look at individual patterns within the group: the distribution of responses in the verb condition looks highly similar between learners and natives, although there is a tendency toward more positivities without preceding negativities and less biphasic responses in the learners. The fact that basically none of the learners show any sensitivity to gender violations assures us that the null effect in the grand mean analysis was not due to a cancelation by different patterns.

Predictors of P600 effect magnitude in the learner group

To investigate which factors lead to a higher degree of native-likeness in L2 learners, we performed a multiple regression analysis (e.g., Baayen, 2008), to investigate the possible influence of age of acquisition, length of residence, L2 proficiency (as measured by the C-test), offline gender knowledge (as measured by the gender assignment task), and L2 use (composite score) on the P600. We took magnitude of the P600 as a measure of native-likeness, since the previous section revealed that this is the most reliable effect in the native group. The average amplitude of the difference wave (incorrect minus correct), calculated in the 600–1200 ms window collapsing middle frontal and all temporal, parietal, and occipital ROIs, was used as the dependent measure in the regression model. Because of skewed distributions, age of acquisition, and length of residence were log-transformed, and L2 proficiency, gender knowledge and L2

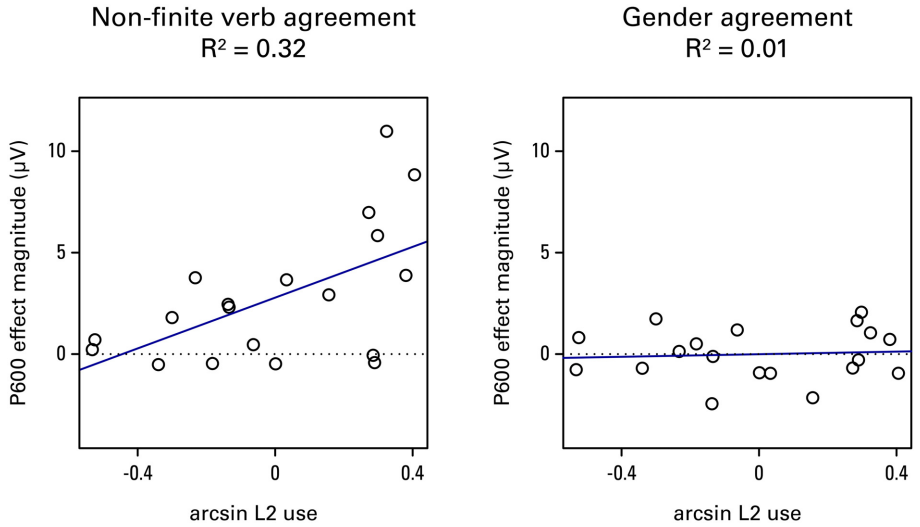


FIGURE 7 | The percentage of use of the L2 in daily life predicts P600 magnitude for non-finite verb agreement violations, but not gender agreement violations.

Table 3 | Correlation matrix for the dependent measure and the participant characteristics variables used in the regression model.

	P600 magnitude	Log age of acquisition	Log length of residence	Arcsin proficiency	Arcsin gender knowledge	Arcsin L2 use
P600 magnitude	–					
Log age of acquisition	−0.083	–				
Log length of residence	−0.106	−0.147	–			
Arcsin proficiency	0.140	−0.327	0.230	–		
Arcsin gender knowledge	0.134	−0.416	0.552*	0.424	–	
Arcsin L2 use	0.486*	−0.388	0.413	0.293	0.518*	–

Asterisk indicates significance of $p < 0.05$.

use were arcsine transformed prior to entry into the model. Additionally all predictor variables were centered at their mean. The correlation matrix for the dependent measure and the participant characteristics variables can be found in **Table 3**. Examining **Table 3** we see that length of residence shows a significant positive correlation with gender knowledge (i.e., the ability to assign gender offline), $r_{(17)} = 0.55$, $p = 0.014$, with longer length of residence being associated with better gender knowledge. However, there is no relation between length of residence and the magnitude of the P600 (i.e., the ability to process grammatical structures efficiently online), $r_{(17)} = -0.11$, $p = 0.665$. L2 use positively correlates with both gender knowledge and P600 magnitude, $r_{(17)} = 0.52$, $p = 0.023$ and $r_{(17)} = 0.49$, $p = 0.035$, respectively, with a higher amount of L2 use being associated with better gender knowledge as well as larger P600 magnitudes.

In addition to the participant characteristics variables, structure and modality were tested as predictors in the model. The significance of predictors was evaluated by means of the *t*-test for the coefficients, in addition to model comparison using AIC (Akaike Information Criterion; Akaike, 1974). **Table 4** shows the best linear multiple regression model (explained variance: 33.7%). This model shows that the structure being gender has a negative impact ($\beta = -2.79$, $t = -4.41$), and L2 use has a positive impact ($\beta = 3.29$, $t = 3.07$) on P600 effect magnitude. The other predictors (i.e., modality, age of acquisition, length of residence, proficiency, and gender knowledge) did not reach significance by themselves or in interaction with any other variables and were therefore not included in the model. Finally, the model additionally shows an interaction between the structure being gender and L2 use ($\beta = -5.94$, $t = -2.78$). This effect is plotted in **Figure 7**. There appears to be a significant effect of L2 use on the P600 for non-finite verb agreement violations, $R^2 = 0.32$, $F_{(1,\ 17)} = 8.08$, $p = 0.011$, but no significant effect for gender agreement violations, $R^2 = 0.01$, $F_{(1,\ 17)} = 0.01$, $p = 0.756$.

Table 4 | Linear multiple regression model predicting P600 effect magnitude in learners.

Predictor	Estimate	*SE*	*t*-value	*p*-value
Intercept	1.388	0.316	4.390	<0.001
StructureIsGender	−2.789	0.632	−4.410	<0.001
L2use	3.288	1.070	3.074	0.003
StructureIsGender*L2use	−5.939	2.140	−2.776	0.007

No other significant interactions with structure or modality were found.

DISCUSSION

Using the P600 as a measure of native-likeness, we tested whether sufficiently proficient late L2 learners can show native-like syntactic processing, even if (1) gender marking in the L1 is implemented differently and (2) the L2 gender system is opaque. We investigated the ERP responses of native speakers and Romance learners of Dutch to anomalies in constructions that are relatively easy to acquire (i.e., non-finite verbs) and those that have been shown to be more difficult (i.e., gender). In addition, we varied the modality in which the stimuli were presented, in order to investigate whether visual presentation might contribute to the lack of sensitivity to gender in the Romance group reported in previous research (Sabourin and Stowe, 2008). The non-finite verb violations elicited a biphasic N400-P600 effect in both native speakers and second language learners. However, in contrast to the native speakers, the learners only showed evidence of an N400 in the auditory and not the visual condition, although the statistical support for this difference is weak[3]. Also, the amplitude of the N400 effect was somewhat smaller than in the natives. For the gender violations, we found a clear P600 in natives, but not in L2 learners.

The effects of modality were quite subtle. We had hypothesized that increased processing demands in the visual modality might interfere with immersed learners' responses to grammatical violations and that they might show more native-like responses in the auditory modality. This hypothesis receives some support; the modulation of the N400 effect in non-finite verb violations in learners was in the hypothesized direction, with a native-like effect in the auditory but not the visual modality. However, for gender agreement learners failed to show sensitivity, regardless of the modality. Thus, the suggestion that the difference between Loerts' (2012) results for Polish speakers on the one hand, and Sabourin and Stowe's (2008) results and our current results for Romance speakers on the other, cannot be attributed to the difference in modalities.

In contrast to the modality effects, violation effects and group differences therein were robust. Before accepting the group patterns, it is important to examine the role of individual differences. A biphasic pattern may reflect the summation of single effects originating in two different groups of participants (Osterhout, 1997; Nieuwland and Van Berkum, 2008; Tanner and Van Hell, 2014; Tanner et al., 2014). Even more crucial for the current experiment, the absence of an effect in the L2 group may be due to variability, with some individuals showing the pattern found in native speakers, while others show no effect or even an opposing effect (Foucart and Frenck-Mestre, 2011). Inspection of individual differences for the gender violations confirmed that the grand average ERP patterns we report are representative of the majority of the individuals in each group. In contrast to natives, who consistently showed large P600 effects (**Figure 6**, bottom left panel), learners consistently failed to demonstrate any form of sensitivity to gender violations (**Figure 6**, bottom right panel). This result was confirmed by the fact that none of the participant characteristics we tested (increased proficiency or gender knowledge, earlier age of acquisition, longer length of residence or high percentage L2 use) was associated with a larger P600. In this sense, the current experiment replicates the pattern found by Sabourin and Stowe (2008); even highly proficient Romance learners of Dutch appear to have persistent difficulties in learning to use Dutch gender.

Turning to the non-finite verb violations, examination of the native speakers confirms that the biphasic pattern N400/P600 seen in this group is present in the majority of the individual participants (see **Figure 6**, top left panel). This biphasic effect in response to non-finite verb violations in natives has been found before (Kutas and Hillyard, 1983; Sabourin and Stowe, 2008; Loerts, 2012). As can be seen in **Figure 6** (top right panel), many learners' responses were within the native range, showing evidence of the biphasic pattern, although this is primarily evident for the auditorily presented materials. Some individuals are less native-like; for this structure the P600 effect magnitude in the L2 group was found to be modulated by the percentage of use of the L2 in daily life. Use is not the only important factor for native-like attainment of syntax processing however; even the learners with the highest amount of daily practice in an immersed setting still show persistent problems with gender agreement.

Despite their failure to show native-like gender processing, the evidence suggests that the Romance learners are highly proficient. In addition to the off-line measures of proficiency (C-test and gender assignment) and online accuracy at ungrammaticality detection, which are within native range for a number of the participants, the evidence from the biphasic N400-P600 pattern provides a strong argument for high proficiency. Finding early ERP effects in response to grammatical violations like the N400 seen here is unusual in L2 research. Although both Loerts (2012) and Sabourin and Stowe (2008) found evidence of a biphasic pattern for their native groups, neither found the N400 in their L2 learner groups. According to Steinhauer et al. (2009), biphasic patterns are one of the latest stages of morpho-syntactic proficiency in late L2 acquisition. The fact that our learners were able to reach this stage for non-finite verb agreement, but that they cannot get past the initial stage of not showing any brain response differences for correct vs. incorrect use of gender agreement provides strong support for the difficulty of the acquisition of this element in Dutch L2 acquisition. This highlights the complexity of acquisition of the Dutch gender system, even by learners with a gender system in their L1. Furthermore, it emphasizes the fact that language learning aptitude is not an all or none phenomenon, but may vary widely between constructions.

Our results further illustrate the large discrepancy between online and offline processing measures in L2 acquisition research. Both the behavioral results of the gender assignment task and the

[3]We want to remind the reader that the modality effect in the non-finite verb condition should be interpreted with some caution. Unlike the main effects we report throughout the rest of the discussion (which are based on 24 and 48 items per condition for verb and gender, respectively), the marginally significant interaction we followed up on here is based on 12 items per condition only, which is relatively few for an ERP study. However, if we do not follow up on this interaction the main effect of correctness remains, suggesting that learners are like natives. We felt this claim would be too strong, and therefore discuss the follow-up analysis, despite the statistical concerns.

sentence-final grammaticality judgments during the ERP recordings for gender violations indicate moderate to good knowledge of Dutch grammatical gender in the learner group. Yet, we observed a complete lack of response to these violations in the ERP signal. This reveals a discrepancy between offline knowledge of grammatical gender concord and the use of agreement knowledge during online processing. The lack of a significant relation between the magnitude of the P600 responses to gender violations and the score on the gender assignment task rules out the possibility that only learners with better offline performance are able to show online effects. The behavioral difference between the visual and the auditory modality, with performance being slightly worse for grammaticality judgments in the auditory modality, was also not reflected in the ERP signal for gender violations. These results illustrate that second language learners can develop successful strategies to cope with gender processing difficulties. These alternative routes, however, apparently take more time and are qualitatively different from what we observe in online native processing.

The results of the current study leave us with a puzzle; why do Romance learners of Dutch show such persistent problems with gender processing? Our results confirm that gender is difficult to process for late Romance learners of Dutch, compared with the results of studies targeting other languages. We replicated Sabourin and Stowe's (2008) findings, in the sense that our Romance learners likewise did not show native-like responses to gender violations, regardless of modality, although they showed responses to non-finite verbs that were close to the native model[4]. The factors most commonly suggested in the literature as to why gender or other forms of grammatical processing might be problematic do not appear to explain these results. Proficiency clearly plays some role in native-likeness in general (Steinhauer et al., 2006; McLaughlin et al., 2010), but as we argue above, our learners were quite proficient, certainly comparable to those in other studies in which learners have shown P600 effects for gender (Tokowicz and MacWhinney, 2005; Frenck-Mestre et al., 2009; Gillon Dowens et al., 2010, 2011; Foucart and Frenck-Mestre, 2011, 2012; Loerts, 2012). Also, our proficiency measure does not correlate with the magnitude of the ERPs.

Other potential explanatory factors involve the language experience of the learner, such as age of acquisition (Weber-Fox and Neville, 1996; Kotz et al., 2008) and exposure to and use of the L2 (Gardner et al., 1997; Flege et al., 1999; Dörnyei, 2005; Tanner et al., 2014). It is true that the studies reported by Frenck-Mestre and colleagues have generally tested earlier learners (with onset of acquisition in their teens rather than twenties and later). However, other studies have demonstrated native-like gender processing even for relatively late learners (Tokowicz and MacWhinney, 2005; Gillon Dowens et al., 2010). Furthermore, in the current study we did not even find a trend toward better performance for younger learners, making it again unlikely that this is the (only) decisive factor for native-likeness. The amount of L2 use also failed to explain the failure of the Romance learners to show online sensitivity to gender, even though, as our own results show, this can be important for native-likeness for other aspects of grammatical processing, like verb agreement. Length of residence, which impacts overall exposure, also showed no correlation with sensitivity to gender.

Failure to achieve native-like processing has also been linked to dissimilarity between L1 and L2 (Tokowicz and MacWhinney, 2005; Sabourin and Stowe, 2008; Foucart and Frenck-Mestre, 2011), as well as characteristics of the target language (Sabourin and Stowe, 2008; Loerts, 2012). Following this line of argumentation, Dutch and Romance languages may simply be too different from each other, which, combined with the fact that the Dutch gender system is relatively opaque, results in a very difficult challenge for native-like attainment. The lack of transparency of the Dutch gender system might explain why our Romance learners failed to show native-like processing for this characteristic of the language, as opposed to the much more transparent non-finite verb manipulation. For gender, previous research has shown that native-like processing is possible even in constructions with competition from an L1 gender system when a relatively transparent target gender system is to be acquired in L2 (Frenck-Mestre et al., 2009; Foucart and Frenck-Mestre, 2011; Gillon Dowens et al., 2011). In contrast, Loerts' study suggests that an opaque system is more difficult to acquire, since only her most proficient learners are able to show P600 effects, which are additionally somewhat smaller in amplitude compared to the natives. It remains an open question as to why, in contrast to Loerts (2012), even the most proficient learners in the current study did not show a P600. More research is needed to determine whether characteristics of the L1 or other (confounding) factors are at play in determining which individuals overcome the challenge of an opaque gender system.

One final point we would like to make is that, although we did not find extensive effects of stimulus modality, this factor is nevertheless of importance. As we noted, the early responses to ungrammaticality like the N400 in the biphasic response seen here are not generally found in late L2 learners, which has been taken as a sign of lack of native-likeness. It is possible that they have been missed due to the use of visual materials, since this effect was only seen in the auditory modality. Although we saw no effects on the amplitude of the P600 effect, certain populations may be affected more than others. Learners who do not share the same writing system in their L1 and L2, for instance, might have more difficulty automatizing their usage of the new alphabet (Koda, 1999; Wang et al., 2003). For these learners, the use of auditory materials might be a crucial prerequisite to obtain an accurate measure of their abilities. On the other hand, those whose learning has taken place with an emphasis on written materials may show less response when auditory materials are used. Given the large diversity of L2 speaker populations with respect to typological distance (both with respect to grammar and writing systems) and type of learning environment (immersion vs. classroom), it is important to be aware that the testing modality might influence the results, both in offline and online tests.

[4]One of the reviewers points out that having twice as many violation sentences in the gender condition than the non-finite verb condition, might be problematic, since less common stimulus types may elicit a P3 response (see Coulson et al., 1998; Hahne and Friederici, 1999). However, the difference waves shown overlaid in **Figure 5** show that there is no difference in P600 effect magnitude between gender and non-finite verbs in the natives.

In conclusion, we can say that online grammatical gender processing is particularly difficult for Romance learners of Dutch, even at high levels of proficiency and with large amounts of L2 exposure and use in a natural setting, and regardless of testing modality. In contrast, responses highly similar to the native model are possible for a more regular and transparent structure (non-finite verbs), for which responses are modulated by both testing modality and L2 use. In contrast, the problems with gender are persistent and not affected by these factors, demonstrating the complexity of (late) L2 acquisition of the opaque Dutch gender system.

ACKNOWLEDGMENTS

We are very grateful for comments and suggestions of the reviewers. This research was supported by the Netherlands Organization for Scientific Research (NWO) under grant 016.104.602, awarded to the fifth author. We thank Hanneke Loerts, Sanne Berends and Bregtje Seton for sharing their auditory materials, and the participants for their kind cooperation. Additionally we thank our colleagues at the NeuroImaging Center Groningen for technical support, particularly Peter Albronda.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01072/abstract

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511801686

Baayen, R. H., Piepenbrock, R., and Gullikers, L. (1995). *The CELEX Lexical Database [CD-ROM]*. Philadelphia, PA: Linguistics Data Consortium, University of Pennsylvania.

Balconi, M., and Pozzoli, U. (2005). Comprehending semantic and grammatical violations in Italian. N400 and P600 comparison with visual and auditory stimuli. *J. Psycholinguist. Res.* 34, 71–98. doi: 10.1007/s10936-005-3633-6

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. R package version 1.1-6. Available online at: http://CRAN.R-project.org/package=lme4

Bates, E., and MacWhinney, B. (1987). "Competition, variation, and language learning," in *Mechanisms of Language Acquisition, The 20th Annual Carnegie Symposium on Cognition* (Hillsdale, NJ: Lawrence Erlbaum Associates), 157–193.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300.

Bernard, M. L., Chaparro, B. S., and Russell, M. (2001). Examining automatic text presentation for small screens. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 45, 637–639. doi: 10.1177/154193120104500613

Blom, E., Polišenská, D., and Weerman, F. (2008). Articles, adjectives and age of onset: the acquisition of Dutch grammatical gender. *Second Lang. Res.* 24, 297–331. doi: 10.1177/0267658308090183

Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Res. Rev.* 59, 55–73. doi: 10.1016/j.brainresrev.2008.05.003

Bruhn de Garavito, J., and White, L. (2000). "Second language acquisition of Spanish DPs: the status of grammatical features," in *BUCLD 24: Proceedings from the 24th Annual Boston University Conference on Language Development*, eds S. C. Howell, S. Fish, and T. Keith-Lucas (Somerville, MA: Cascadilla), 164–175.

Burkhardt, P. (2007). The P600 reflects cost of new information in discourse memory. *Neuroreport* 18, 1851–1854. doi: 10.1097/WNR.0b013e3282f1a999

Camblin, C. C., Ledoux, K., Boudewyn, M., Gordon, P. C., and Swaab, T. Y. (2007). Processing new and repeated names: effects of coreference on repetition priming with speech and fast RSVP. *Brain Res.* 1146, 172–184. doi: 10.1016/j.brainres.2006.07.033

Clahsen, H., and Felser, C. (2006). Grammatical processing in language learners. *Appl. Psycholinguist.* 27, 3–42. doi: 10.1017/S0142716406060024

Corbett, G. (1991). *Gender*. Cambridge: Cambridge University Press.

Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: event-related brain response to morphosyntactic violations. *Lang. Cogn. Process.* 13, 21–58. doi: 10.1080/016909698386582

Davidson, D. J., and Indefrey, P. (2009). An event-related potential study on changes of violation and error responses during morphosyntactic learning. *J. Cogn. Neurosci.* 21, 433–446. doi: 10.1162/jocn.2008.21031

Dimitrova, D. V., Stowe, L. A., Redeker, G., and Hoeks, J. C. (2012). Less is not more: neural responses to missing and superfluous accents in context. *J. Cogn. Neurosci.* 24, 2400–2418. doi: 10.1162/jocn_a_00302

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annu. Rev. Appl. Linguist.* 30, 149–166. doi: 10.1017/S026719051000005X

Flege, J. E., Yeni-Komshian, G. H., and Liu, S. (1999). Age constraints on second-language acquisition. *J. Mem. Lang.* 41, 78–104. doi: 10.1006/jmla.1999.2638

Foucart, A., and Frenck-Mestre, C. (2011). Grammatical gender processing in L2: electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingual. Lang. Cogn.* 14, 379–399. doi: 10.1017/S136672891000012X

Foucart, A., and Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *J. Mem. Lang.* 66, 226–248. doi: 10.1016/j.jml.2011.07.007

Franceschina, F. (2005). *Fossilized Second Language Grammars: The Acquisition of Grammatical Gender*. Amsterdam: John Benjamins. doi: 10.1075/lald.38

Frenck-Mestre, C., Foucart, A., Carrasco, H., and Herschensohn, J. (2009). Processing of grammatical gender in French as a first and second language evidence from ERPs. *Eurosla Yearb.* 9, 76–106. doi: 10.1075/eurosla.9.06fre

Friederici, A. D., Wang, Y., Herrmann, C. S., Maess, B., and Oertel, U. (2000). Localization of early syntactic processes in frontal and temporal cortical areas: a magnetoencephalographic study. *Hum. Brain Mapp.* 11, 1–11. doi: 10.1002/1097-0193(200009)11:1%3C1::AID-HBM10%3E3.0.CO;2-B

Frisch, S., Kotz, S. A., von Cramon, D. Y., and Friederici, A. D. (2003). Why the P600 is not just a P300: the role of the basal ganglia. *Clin. Neurophysiol.* 114, 336–340. doi: 10.1016/S1388-2457(02)00366-8

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: true issues and false trails. *Psychol. Bull.* 123, 71–99. doi: 10.1037/0033-2909.123.1.71

Gardner, R. C., Tremblay, P. F., and Masgoret, A. (1997). Towards a full model of second language learning: an empirical investigation. *Mod. Lang. J.* 81, 344–362. doi: 10.1111/j.1540-4781.1997.tb05495.x

Gillon Dowens, M., Guo, T., Guo, J., Barber, H., and Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish—evidence from event related potentials. *Neuropsychologia* 49, 1651–1659. doi: 10.1016/j.neuropsychologia.2011.02.034

Gillon Dowens, M., Vergara, M., Barber, H. A., and Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *J. Cogn. Neurosci.* 22, 1870–1887. doi: 10.1162/jocn.2009.21304

Gratton, G., and Coles, M. H. (1989). Generalization and evaluation of eye-movement correction procedures. *J. Psychophysiol.* 3, 14–16.

Grüter, T., Lew-Williams, C., and Fernald, A. (2012). Grammatical gender in L2: a production or a real-time processing problem? *Second Lang. Res.* 28, 191–215. doi: 10.1177/0267658312437990

Gunter, T. C., Stowe, L. A., and Mulder, G. (1997). When syntax meets semantics. *Psychophysiology* 34, 660–676. doi: 10.1111/j.1469-8986.1997.tb02142.x

Hagoort, P., Brown, C., and Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Lang. Cogn. Process.* 8, 439–483. doi: 10.1080/01690969308407585

Hagoort, P., and Brown, C. M. (2000). ERP effects of listening to speech compared to reading: the P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia* 38, 1531–1549. doi: 10.1016/S0028-3932(00)00053-1

Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *J. Psycholinguist. Res.* 30, 251–266. doi: 10.1023/A:1010490917575

Hahne, A., and Friederici, A. (1999). Electrophysiological evidence for two steps in syntactic analysis: early automatic and late controlled processes. *J. Cogn. Neurosci.* 11, 194–205. doi: 10.1162/089892999563328

Hawkins, R. (2001). The theoretical significance of Universal Grammar in second language acquisition. *Second Lang. Res.* 17, 345–367. doi: 10.1191/026765801681495868

Hawkins, R., and Chan, C. Y. H. (1997). The partial availability of Universal Grammar in second language acquisition: the "failed functional features hypothesis." *Second Lang. Res.* 13, 187–226. doi: 10.1191/026765897671476153

Homae, F., Hashimoto, R., Nakajima, K., Miyashita, Y., and Sakai, K. L. (2002). From perception to sentence comprehension: the convergence of auditory and visual information of language in the left inferior frontal cortex. *Neuroimage* 16, 883–900. doi: 10.1006/nimg.2002.1138

Hopp, H. (2010). Ultimate attainment in L2 inflection: performance similarities between non-native and native speakers. *Lingua* 120, 901–931. doi: 10.1016/j.lingua.2009.06.004

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: relations between lexical and syntactic variability. *Second Lang. Res.* 29, 33–56. doi: 10.1177/0267658312461803

Keijzer, M. (2007). *Last in First Out? An Investigation of the Regression Hypothesis in Dutch Emigrants in Anglophone Canada*. Ph.D. dissertation, Vrije Universiteit Amsterdam, Netherlands.

Kluender, R., and Kutas, M. (1993). Bridging the gap: evidence from ERPs on the processing of unbounded dependencies. *J. Cogn. Neurosci.* 5, 196–214. doi: 10.1162/jocn.1993.5.2.196

Koda, K. (1996). L2 word recognition research: a critical review. *Mod. Lang. J.* 80, 450–460. doi: 10.1111/j.1540-4781.1996.tb05465.x

Koda, K. (1999). Development of L2 intraword orthographic sensitivity and decoding skills. *Mod. Lang. J.* 83, 51–64. doi: 10.1111/0026-7902.00005

Kotz, S. A., Holcomb, P. J., and Osterhout, L. (2008). ERPs reveal comparable syntactic sentence processing in native and non-native readers of English. *Acta psychol.* 128, 514–527. doi: 10.1016/j.actpsy.2007.10.003

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., and Hillyard, S. A. (1983). Event–related brain potentials to grammatical errors and semantic anomalies. *Mem. Cognit.* 11, 539–550. doi: 10.3758/BF03196991

Lawrence, M. A. (2013). *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.2-2. Available online at: http://CRAN.R-project.org/package=ez

Loerts, H. (2012). *Uncommon Gender: Eyes and Brains, Native and Second Language Learners, and Grammatical Gender*. Ph.D. dissertation, Rijksuniversiteit Groningen, Grodil.

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., et al. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Lang. Learn.* 60, 123–150. doi: 10.1111/j.1467-9922.2010.00604.x

Molinaro, N., Barber, H. A., and Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex* 47, 908–930. doi: 10.1016/j.cortex.2011.02.019

Müller, H. M., King, J. W., and Kutas, M. (1997). Event-related potentials elicited by spoken relative clauses. *Cogn. Brain Res.* 5, 193–203. doi: 10.1016/S0926-6410(96)00070-5

Münte, T. F., Heinze, H. J., and Mangun, G. R. (1993). Dissociation of brain activity related to syntactic and semantic aspects of language. *J. Cogn. Neurosci.* 5, 335–344. doi: 10.1162/jocn.1993.5.3.335

Nieuwland, M. S., and Van Berkum, J. J. (2008). The interplay between semantic and referential aspects of anaphoric noun phrase resolution: evidence from ERPs. *Brain Lang.* 106, 119–131. doi: 10.1016/j.bandl.2008.05.001

Osterhout, L. (1997). On the brain response to syntactic anomalies: manipulations of word position and word class reveal individual differences. *Brain Lang.* 59, 494–522. doi: 10.1006/brln.1997.1793

Osterhout, L., and Hagoort, P. (1999). A superficial resemblance does not necessarily mean you are part of the family: counterarguments to Coulson, King and Kutas (1998) in the P600/SPS-P300 debate. *Lang. Cogn. Process.* 14, 1–14. doi: 10.1080/016909699386356

Osterhout, L., and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* 31, 785–806. doi: 10.1016/0749-596X(92)90039-Z

Perfetti, C. A., Zhang, S., and Berent, I. (1992). Reading in English and Chinese: evidence for a "universal" phonological principle. *Adv. Psychol.* 94, 227–248. doi: 10.1016/S.0166-4115(08)62798-3

R Core Team. (2014). *R: A Language And Environment For Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rossi, S., Gugler, M. F., Friederici, A. D., and Hahne, A. (2006). The impact of proficiency on syntactic second-language processing of German and Italian: evidence from event-related potentials. *J. Cogn. Neurosci.* 18, 2030–2048. doi: 10.1162/jocn.2006.18.12.2030

Sabourin, L. (2003). *Grammatical Gender and Second Language Processing: An ERP Study*. Ph.D. dissertation, Groningen, Grodil.

Sabourin, L., and Stowe, L. A. (2008). Second language processing: when are first and second languages processed similarly? *Second Lang. Res.* 24, 397–430. doi: 10.1177/0267658308090186

Schneider, W., Eschman, A., and Zuccolotto, A. (2002a). *E-Prime User's Guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Schneider, W., Eschman, A., and Zuccolotto, A. (2002b). *E-Prime Reference Guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Schwartz, B. D., and Sprouse, R. A. (1994). "Word order and nominative case in nonnative language acquisition: a longitudinal study of (L1 Turkish) German Interlanguage," in *Language Acquisition Studies in Generative Grammar: Papers in Honor of Kenneth Wexler from the 1991 GLOW Workshops*, eds T. Hoekstra and B. D. Schwartz (Philadelphia, PA: John Benjamins), 317–368.

Schwartz, B. D., and Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access model. *Second Lang. Res.* 12, 40–72. doi: 10.1177/026765839601200103

Steinhauer, K., White, E., Cornell, S., Genesee, F., and White, L. (2006). The neural dynamics of second language acquisition: evidence from event-related potentials. *J. Cogn. Neurosci.* (Suppl. 99).

Steinhauer, K., White, E. J., and Drury, J. E. (2009). Temporal dynamics of late second language acquisition: evidence from event-related brain potentials. *Second Lang. Res.* 25, 13–41. doi: 10.1177/0267658308098995

Stowe, L. A. (1991). Ambiguity resolution: behavioral evidence for a delay. *Proceedings of the Thirteenth Annual Meeting of the Cognitive Science Association* (Hillsdale, NJ: Lawrence Erlbaum Associates), 257–262.

Tanner, D., Inoue, K., and Osterhout, L. (2014). Brain-based individual differences in online L2 grammatical comprehension. *Bilingual. Lang. Cogn.* 17, 277–293. doi: 10.1017/S1366728913000370

Tanner, D., McLaughlin, J., Herschensohn, J., and Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingual. Lang. Cogn.* 16, 367–382. doi: 10.1017/S1366728912000302

Tanner, D., and Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia* 56, 289–301. doi: 10.1016/j.neuropsychologia.2014.02.002

Tokowicz, N., and MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: an event-related potential investigation. *Stud. Second Lang. Acquis.* 27, 173–204. doi: 10.1017/S0272263105050102

van Berkum, J. (1996). *The Psycholinguistics of Grammatical Gender: Studies in Language Comprehension and Production*. Ph.D. dissertation, Max Planck Institute for Psycholinguistics. Nijmegen, Nijmegen University press.

Wang, M., Koda, K., and Perfetti, C. A. (2003). Alphabetic and nonalphabetic L1 effects in English word identification: a comparison of Korean and Chinese English L2 learners. *Cognition* 87, 129–149. doi: 10.1016/s0010-0277(02)00232-9

Weber-Fox, C. M., and Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence

in bilingual speakers. *J. Cogn. Neurosci.* 8, 231–256. doi: 10.1162/jocn.1996.8.3.231

Weber-Fox, C. M., and Neville, H. J. (1999). "Functional neural subsystems are differentially affected by delays in second language immersion: ERP and behavioral evidence in bilinguals," in *Second Language Acquisition and the Critical Period Hypothesis*, ed D. Birdsong (Mahwah, NJ: Erlbaum), 23–38.

White, L. (1989). *Universal Grammar and Second Language Acquisition,* Vol. 1. Amsterdam: John Benjamins Publishing.

White, L. (2007). "Some puzzling features of L2 features," in *The Role of Features in Second Language Acquisition*, eds J. Liceras, H. Zobl, and H. Goodluck (Mahwah, NJ: Erlbaum), 305–330.

White, L., Valenzuela, E., Kozlowska-Macgregor, M., and Leung, I. (2004). Gender and number agreement in nonnative Spanish. *Appl. Psycholinguist.* 25, 105–133. doi: 10.1017/S0142716404001067

White, L., Valenzuela, E., Kozlowska-Macgregor, M., Leung, I., and Ayed, H. B. (2001). "The status of abstract features in Interlanguage: gender and number in L2 Spanish," in *BUCLD 25 Proceedings* (Somerville, MA: Cascadilla Press), 792–802.

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: 16 May 2014; accepted: 06 September 2014; published online: 25 September 2014.

Citation: Meulman N, Stowe LA, Sprenger SA, Bresser M and Schmid MS (2014) An ERP study on L2 syntax processing: When do learners fail? Front. Psychol. ***5****:1072. doi: 10.3389/fpsyg.2014.01072*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 09 February 2015
doi: 10.3389/fpsyg.2015.00077

Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English

***Cecile De Cat*[1]*, Ekaterini Klepousniotou*[2] *and R. Harald Baayen*[3]**

[1] *Department of Linguistics and Phonetics, University of Leeds, Leeds, UK*
[2] *School of Psychology, University of Leeds, Leeds, UK*
[3] *Quantitative Linguistics Lab, Department of Linguistics, Eberhard Karls University Tübingen, Tübingen, Germany*

Edited by:
Vicky Chondrogianni, University of Edinburgh, UK

Reviewed by:
Niels O. Schiller, University of Leiden, Netherlands
Giorgio Arcara, IRCCS, Fondazione Ospedale San Camillo, Italy
Tanja Rinker, University of Konstanz, Germany

***Correspondence:**
Cecile De Cat, Department of Linguistics and Phonetics, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK
e-mail: c.decat@leeds.ac.uk

The processing of English noun-noun compounds (NNCs) was investigated to identify the extent and nature of differences between the performance of native speakers of English and advanced Spanish and German non-native speakers of English. The study sought to establish whether the word order of the equivalent structure in the non-native speakers' mothertongue (L1) had an influence on their processing of NNCs in their second language (L2), and whether this influence was due to differences in grammatical representation (i.e., incomplete acquisition of the relevant structure) or processing effects. Two mask-primed lexical decision experiments were conducted in which compounds were presented with their constituent nouns in licit vs. reversed order. The first experiment used a speeded lexical decision task with reaction time registration, and the second a delayed lexical decision task with EEG registration. There were no significant group differences in accuracy in the licit word order condition, suggesting that the grammatical representation had been fully acquired by the non-native speakers. However, the Spanish speakers made slightly more errors with the reversed order and had longer response times, suggesting an L1 interference effect (as the reverse order matches the licit word order in Spanish). The EEG data, analyzed with generalized additive mixed models, further supported this hypothesis. The EEG waveform of the non-native speakers was characterized by a slightly later onset N400 in the violation condition (reversed constituent order). Compound frequency predicted the amplitude of the EEG signal for the licit word order for native speakers, but for the reversed constituent order for Spanish speakers—the licit order in their L1—supporting the hypothesis that Spanish speakers are affected by interferences from their L1. The pattern of results for the German speakers in the violation condition suggested a strong conflict arising due to licit constituents being presented in an order that conflicts with the expected order in both their L1 and L2.

Keywords: compounds, second language, word order, ERP, frequency effects, generalized additive mixed models

1. INTRODUCTION

Noun-noun compounds are entities consisting in two nouns united by a semantic relation (Gagné and Spalding, 2014) that is not overtly expressed. Endocentric compounds contain a head element (*dust* in 1) whose lexical category and interpretive features are inherited by the compound and contribute the core of its meaning (e.g., a kind of dust). The other element acts as a modifier of that head.

(1) moon dust ("dust from the moon" / "dust made of moon" / "dust with moon-like properties")

Compounds have been extensively studied in the past 40 years from a myriad of viewpoints (Libben and Jarema, 2006; Lieber and Štekauer, 2009; Semenza and Luzzatti, 2014). A key concern has been whether the processing of compounds consists of retrieving entities listed in the mind (Butterworth, 1983) or requires decomposition into constituents listed separately (Semenza et al., 1997; Libben, 1998). Dual-route theories contend that the two processes (i.e., a whole-word and a parsing procedure) exist side by side (Sandra, 1990). It is now widely accepted that both constituents are activated during processing, at least in non-lexicalised compounds (Jarema, 2006; Zhang et al., 2012; MacGregor and Shtyrov, 2013). Noun-noun compounds have also been shown to be processed differently to non-compounds of similar morphological complexity and length, with compounds yielding longer reaction times and different electrophysiological correlates (El Yagoubi et al., 2008).

Here we focus on endocentric noun-noun compounds (henceforth NNCs), which have been argued to embody an underlying structure (Libben, 2006): their structure is hierarchical, involving the (possibly recursive) subordination of a modifier to a grammatical head (or a modifier-head compound, as in 2-b).

(2) a. [[[lunch box] lid] stack]
b. [child[amateur [puppet theater]]]

These characteristics suggest that NNCs involve phrasal syntax. Diachronic and synchronic corroborating evidence is provided by (Zipser, 2013): cross-linguistically, (i) the constituent order of compounds reflects the current word order or an earlier word order found in the underlying phrases (e.g., *nut-cracker* shows SOV, the Old English word order); (ii) adjective-noun compounds are not recursive, as predicted by the fact that adjectives do not allow adjective complements; and (iii) recursive compounding is possible only in right-branching phrase structures.

What makes the acquisition of NNCs by non-native speakers particularly interesting to study is that the syntactic properties they exhibit (hierarchical structure, head directionality) are predicted to be acquired very early[1], and their interpretation is essentially a matter of phrasal semantics (which has been shown not to cause persistent difficulty for L2 learners, see Slabakova, 2008). NNCs also appear very early in L1 acquisition (Nicoladis and Yin, 2002; Krott et al., 2010). All this predicts that the processing of NNCs should be relatively unproblematic for advanced learners of English. In particular, L1 word-order effects are not expected: L2ers whose L1 features the opposite word order (i.e., head-first) should not accept English NNCs in reversed order more than L2ers whose L1 order matches that of English. At an advanced level of proficiency, both groups are expected to reject irreversible compounds presented in reversed order:

(3) a. #[[basket] dog] ⟶ uninterpretable as head-last
b. *[basket [dog]] ⟶ head-first order is ungrammatical

Headedness plays a specific role in the processing of NNCs, as shown by research on Italian (which features the two word orders in NNCs): based on a lexical decision task on healthy adults, (El Yagoubi et al., 2008) found clear effects induced by the head, independently of its position in the NNC. Arcara et al. (2014) recently argued that (in Italian) NNCs are decomposed differently, depending on whether they are head-initial or head-final, the latter requiring a higher processing effort when decomposition is elicited. This suggests that in Italian, only head-final compounds are true hierarchical structures (as opposed to lexicalised syntactic units)—see Marelli et al. (2009, 2014). Headedness effects are not distinguishable from position-in-the-string effects in languages such as English. For instance, Jarema et al. (1999) observed no difference in the priming of NNCs by the head or the modifier. This paper takes this line of research further, by investigating whether L1 headedness affects the L2 processing of transparent, irreversible NNCs in very advanced learners of English. In two separate studies, we examined the reaction times and the event-related potentials in response to irreversible NNCs presented in licit vs. reversed word order.

[1]Cross-linguistically, head directionality transfer effects in L2 acquisition have been found to be very short-lived, both in child and in adult learners—see e.g., Haznedar (1997); Unsworth (2005).

Event-related potentials (ERPs) can provide insight into the neural activity associated with the processing of compounds. Functional interpretations can be inferred from the temporal and spatial characteristics of electromagnetic activity, and ERP components can sometimes reveal the engagement of the cognitive processes involved. Our approach in this paper is exploratory (Otten and Rugg, 2005) and will focus on identifying differences in the amplitude of the EEG signal that can be traced back to properties of the participants (such as their language background) and properties of the compounds (such as their frequency of occurrence, and the frequencies of occurrence of their constituents). Inferences based on previously identified ERP components will be drawn in the discussion as appropriate.

Our research questions are: (i) Does non-native processing of NNCs result in different ERP signatures to native processing? (ii) Is non-native processing of NNCs affected by headedness effects from the mother tongue?

We hypothesize that, if very advanced L2 learners are affected by their L1's headedness settings (in spite of the early parameter resetting), the performance of L2ers whose L1 displays the same word order as English (here: German) will be different to that of those whose L1 doesn't (here: Spanish). A significant proportion of erroneous judgements would be taken to indicate a representational deficit (i.e., incomplete acquisition of the target structure). Longer reaction times are expected for both L2 groups, in line with much research on L2 processing (Kroll et al., 2002; Moreno and Kutas, 2005; Clahsen et al., 2013), but significantly longer reaction times in the Spanish group than in the German group would indicate a specific L1 effect. Differences in the processing mechanisms themselves should translate into significantly different ERP signatures across participant groups.

Furthermore, following up on research on compound processing with eye-movement registration (see e.g., Hyönä and Pollatsek, 1998; Pollatsek et al., 2000; Juhasz et al., 2003; Bertram et al., 2004; Kuperman et al., 2008, 2009; Miwa et al., 2014), we expected compound and constituent frequency as covariates to offer enhanced insights into how German and Spanish advanced learners of English differ from native speakers of English when presented with English compounds with constituents presented in the standard as well as in the reversed order. More specifically, we expected that compound frequency, if useful as a predictor, should modulate the EEG amplitude primarily for native speakers, given that less proficient readers have been observed to show decompositional eye-movement patterns (see Kuperman and Van Dyke, 2011, for English). In addition, constituent frequency effects, ubiquitous in the behavioral and eye-tracking literature, should also be detectable. Since compounds with constituents presented in reversed order can only be made sense of by interpreting the constituents, we expected the strongest constituent effects to be present in the reversed condition.

2. MATERIALS AND METHODS

In order to assess whether any L1 headedness effect affects L2ers' processing of NNCs, we carried out two separate studies based on the same task. We registered the accuracy and (i) the timed response or (ii) the electrophysiological response of the brain to visual stimuli presented in the context of a primed lexical decision

task. Stimuli were irreversible NNCs presented in licit (4-a) and reversed order (4-b).

(4) a. coal dust
b. #dust coal

The participant groups differed in mother tongue: English (control group), Spanish or German (experimental groups). Like English, German features productive compounding, with a head-last structure (Meyer, 1993). Whereas in Spanish, compounds are essentially head-first, and not productive (Piera, 1995).

2.1. PARTICIPANTS

Ten native British English speakers, ten native German learners of English and ten native Spanish learners of English took part in each study (i.e., a different group in each study, as detailed in **Table 1**). Non-native participants all had initial second-language exposure after 8 years of age, and all scored above 60% on a cloze test from the Cambridge Certificate in Advanced English. All participants were right-handed based on the Briggs and Nebes inventory (Briggs and Nebes, 1975), had no speech or language difficulties and had normal or corrected-to normal vision. Ethical approval was issued by the School of Psychology, University of Leeds, and informed written consent was obtained from all volunteers.

2.2. STIMULI

Experimental stimuli consisted of prime-target pairs, presented in 4 experimental conditions in a 3 (Group) × 2 (Prime Condition) × 2 (Word Order) design. The prime was either the head (e.g., *dust* in 4) or the modifier (e.g., *coal* in 4) of the intended compound.

The Word Order factor had 2 levels: licit (modifier - head, as in 4-a) or reversed (head—modifier, as in 4-b). All the NNCs were endocentric and featured a transparent modification relationship. All items were tested for irreversibility on an independent group of 30 native speakers[2].

[2]Each compound was presented one by one in licit and reversed order (in randomized order), and participants were asked to rate them by choosing one of the following options: perfectly ok—rare but ok—strange but ok—a bit too strange—very strange—completely bad. These ratings were converted into a numeric score expressed as a percentage. The frequency of the intended compound did not predict its reversibility (Pearson's product-moment correlation: $\rho = 0.14$. $t = 1.5766$, df $= 115$, $p = 0.1176$).

Table 1 | Participant characteristics.

L1	English	German	Spanish
STUDY 1			
Female/Male	7/3	8/2	5/4
Mean age (+ SD)	23;8 (3;10)	23;2 (0;11)	25;4 (6;10)
Mean proficiency (+ SD)		75% (13)	80% (9)
STUDY 2			
Female/Male	4/6	7/3	3/7
Mean age (+ SD)	22;11 (3;3)	26;5 (5;7)	26;11 (5;3)
Mean proficiency (+ SD)		90% (7)	81% (8)

The frequency of the licit compounds and their constituent nouns was estimated from the post-1990 data in Google N-grams. To avoid lexicalisation effects, only compounds with very low frequencies were included (i.e., below 3300—mean = 359.5, compared with a mean of 279,300 for the constituent nouns).

There was a total of 480 test items (based on 120 compounds), of which 234 are included in the present study (as we focus on the Head Prime condition only, and 3 compounds had to be discarded due to spelling inconsistencies between the licit and the reversed word order conditions). All the compounds were with spaced constituents. The items were pseudo-randomized into 8 different orders (assigned randomly to participants) and presented in 4 blocks, with a rest in between[3].

3. STUDY 1: PRIMED LEXICAL DECISION

3.1. PROCEDURE

Participants were tested individually in a single session lasting approximately 20 min. Stimuli were presented visually in light gray text on a black background. Each trial began with a 100 ms mask (#######), after which the prime was presented for 100 ms followed by a second mask (for 50 ms) and the target (for 8000 ms). Participants had to make a lexical decision about the target (as acceptable or not) by pressing (with their right hand) one of two buttons on a hand-held button box (counterbalanced across participants). We recorded accuracy rates and reaction times from the onset of presentation of the target, using E-Prime software.

3.2. RESULTS

Only responses whose reaction times fell between 150 and 5000 ms were included in the analysis, on the assumption that faster responses would not allow sufficient processing time to yield an acceptability judgment, and slower responses are likely to result from conscious processes (0.003% of data were thus excluded). One Spanish participant was excluded due to production of 40% of the responses above the 5000 ms threshold and borderline proficiency given our inclusion criteria.

As seen in **Table 1** (after exclusion of the abovementioned participant), the proficiency of the Spanish group was slightly higher than that of the German group (Wilcoxon rank sum test: $W = 2049133$, $p < 0.0001$).

3.2.1. Accuracy analysis

Table 2 shows that accuracy was very high overall in all groups, and that the predominant type of error was to accept compounds in the reversed order (rather than reject licit compounds).

The responses on the lexical decision task were analyzed with a generalized linear mixed-effect model with a logit link function and binomial variance, using the `lme4` package, version 1.0-4 (Bates et al., 2013) with the "bobyqa" optimizer, using treatment dummy coding for factorial predictors. Only those predictors that contributed to the model fit were retained, as shown in **Table 3**. As a consequence, the frequency covariates, which did not reach

[3]The stimuli and their frequency statistics are given in **Tables A1**, **A2** in the Appendix. The length of the stimuli ranged from 7 to 18 characters (mean:11.9).

Table 2 | Proportions and types of errors across groups in Study 1.

	English	German	Spanish
Accept reversed	5.97	10.29	14.67
Reject licit	3.98	6.77	6.64
Correct	90.06	82.94	78.69

Table 3 | Coefficients of a logistic mixed-effects regression model fitted to the accuracy data of Study 1, and associated statistics.

	Coefficient	Std. Error	Z	p
Intercept	−0.1960	1.0842	−0.1808	0.8565
Word.Order:Reversed	−0.4439	0.2382	−1.8639	0.0623
L1: German	0.0344	0.3751	0.0917	0.9269
L1: Spanish	−0.0110	0.3491	−0.0316	0.9748
Proficiency	3.0819	1.0657	2.8920	0.0038
Word.Order:Reversed by L1: German	−0.1355	0.2743	−0.4938	0.6214
Word.Order:Reversed by L1: Spanish	−0.7004	0.2795	−2.5060	0.0122

The reference level for Word Order is Licit, and for L1: English.

significance, were removed from the model specification. The resulting model provided a substantially improved fit compared to the null-hypothesis model with random intercepts for participant and item only (and with random slopes for word order condition by participant, and participant group by item)[4].

Table 3 indicates that for English speakers, accuracy was higher in the licit word order condition. Furthermore, in the licit word order condition, accuracy levels are comparable in native and non-native speakers, as can also be seen in the left panel of **Figure 1**. In the reversed word order condition, only the Spanish group performed significantly worse than the native speakers. Across groups, greater proficiency afforded higher accuracy, as illustrated in the right panel of **Figure 1**.

3.2.2. Reaction times analysis

An analysis of the response latencies, summarized in **Table 4** and visualized in **Figure 2**, indicated that all groups were faster at rating compounds in the licit word order condition. Only the Spanish group responded significantly slower than the English group. Speed increased with proficiency. The frequency measures did not reach significance nor improve the model fit, and were therefore removed from the final model[5].

4. STUDY 2: EVENT-RELATED POTENTIALS WITH PRIMED LEXICAL DECISION

4.1. PROCEDURE

The stimuli were the same as in Study 1, and participants were subject to the same inclusion criteria (see **Table 1** for details).

Participants were tested individually in a single session lasting approximately one and a half hours. Stimuli were presented visually in light gray text on a black background. Each trial began with the visual presentation of a series of exclamation points (!!!) for 1000 ms, which was a signal for the participant to rest their eyes and blink. After a delay of 100 ms a fixation point (+) was presented for 250 ms to signal that the trial was about to begin and to alert participants that they had to fixate their eyes and avoid eye movements until the next set of exclamation points. A mask (#######) was then presented for 100 ms after which the prime was presented for 100 ms followed by a second mask (#######) for 50 ms and the target for 1000 ms[6]. After a delay of 500 ms a question mark (?) appeared for 2000 ms during which time participants had to make a lexical decision about the target (decide whether or not it was grammatical in English) by pressing one of two buttons on a hand held button box (counterbalanced across participants). Participants were instructed to respond as accurately as possible; accuracy and reaction times (in ms from the onset of the "?") were recorded. (We do however not report on the reaction times below, as they reflected answer to the cue "?" rather than to the stimuli.) After the response (or at the end of 2000 ms if the participant did not respond), there was a delay of 100 ms before the next trial started. The experimental session was preceded by a practice session comprising 20 trials, which was repeated until participants could perform the task and procedure with no errors and no eye movements during the critical period of stimulus presentation (usually one or two practice sessions were required).

The EEG was recorded (Neuroscan Synamps2) from 60 Ag/AgCl electrodes embedded in a cap based on the extended version of the International 10–20 positioning system (Sharbrough et al., 1991) and fitted with QuikCell liquid electrolyte application system (Compumedics Neuroscan). Additional electrodes were placed on the left and right mastoids. Data were recorded using a central reference electrode placed between Cz and CPz. The ground electrode was positioned between Fz and FPz. To monitor eye movements, electro-oculograms (EOGs) were recorded using electrodes positioned at either side of the eyes, and above and below the left eye. At the beginning of the experiment electrode impedances were below 10 kΩ. The analog EEG and EOG recordings were amplified (band pass filter 0.1–100 Hz), and continuously digitized (32-bit) at a sampling frequency of 500 Hz.

Data were processed offline using Neuroscan Edit 4.3 software (Compumedics Neuroscan) and filtered (0.1–40 Hz, 96 dB/Oct, Butterworth zero phase filter), inspected visually and segments contaminated by muscular movement marked as bad. The effect of eye-blink artifacts was minimized by estimating and correcting their contribution to the EEG using a regression procedure which involves calculating an average blink from 32 blinks for each participant, and removing the contribution of the blink from all other channels on a point-by-point basis. Data were epoched between –100 and 1100 ms relative to the onset of the experimental targets and baseline-corrected by subtracting the mean amplitude over the pre-stimulus interval. Epochs were rejected if

[4]**Table A3** in the Appendix gives a summary of the random effects for this model.

[5]**Table A4** in the Appendix gives a summary of the random effects for this model. **Table A5** in the Appendix give the mean reaction times by participant group and condition.

[6]The average visual angle subtended was 5.7°: the stimuli extended approximately 2.8° to the left and right of the center of the screen.

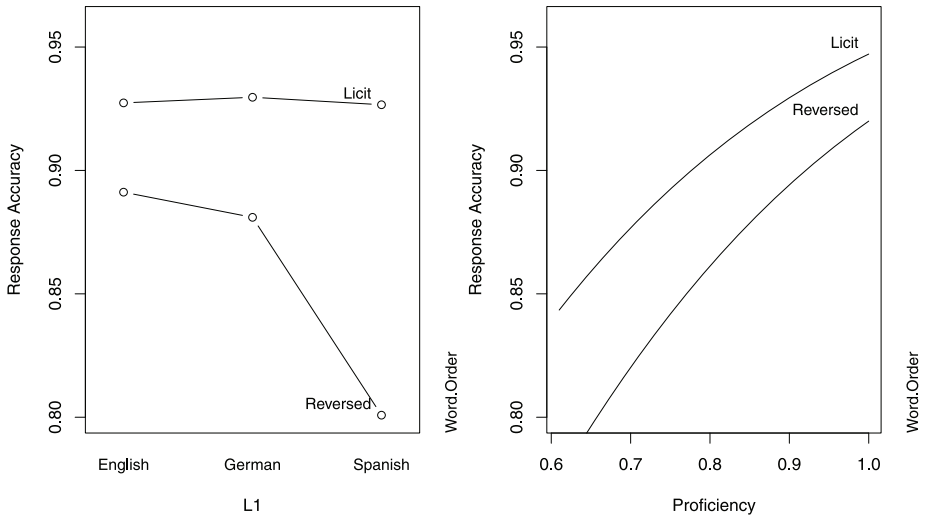


FIGURE 1 | Partial effects of the predictors in the logistic model for response accuracy in Study 1. The left panel is calibrated for the reference levels of Word Order (Licit) and L1 (English), and the right panel is calibrated for median proficiency.

Table 4 | Coefficients of a logistic mixed-effects regression model fitted to the reaction time data.

	Coefficient	Std. Error	*t*-value
(Intercept)	8.2770	0.3100	26.7140
Word.Order:Reversed	0.1230	0.0180	6.7140
L1German	0.0600	0.1020	0.5840
L1Spanish	0.1990	0.0940	2.1160
Proficiency	−1.5620	0.3060	−5.1100

The reference level for Word Order is Licit, and for L1, English. Absolute values of t exceeding 2 are indicative of significance at the 5% level.

participants did not make a response within the allocated time (during presentation of the "?"), or if they made an incorrect response. Subsequently the data was downsampled to 125 Hz. Trial rejection was not done *a priori* but based on the residuals of the modeling, resulting in only 0.7% of discarded data.

4.2. RESULTS

4.2.1. Accuracy analysis

A mixed-effects logistic regression model was fitted to the accuracy data. Results are summarized in **Table 5**, and displayed in **Figure 3**[7]. For English speakers, accuracy did not differ significantly for the licit and reversed word order conditions. For both groups of non-native speakers, accuracy was higher in the Licit Word Order condition, compared with the Reversed Word Order condition. Across groups, greater proficiency afforded higher accuracy.

The main difference in the pattern of results therefore concerns the effect of the word order manipulation, which adversely affected responses for English speakers in the reversed condition in "immediate" lexical decision, but had no consequences for English speakers in the delayed lexical decision task. In addition, when responses are delayed, German speakers pattern together with the Spanish speakers in their response behavior.

4.2.2. ERP analysis

We include for analysis only trials that elicited a correct response. The time window analyzed was limited to 0–800 ms, time-locked to the onset of stimulus presentation[8].

We analyzed the electrophysiological response elicited by the presentation of compound words with the generalized additive mixed model (GAMM, Wood, 2004, 2006; Tremblay and Baayen, 2010; Kryuchkova et al., 2012; Tremblay and Newman, 2015; Baayen, in preparation; Baayen et al., in preparation). Generalized additive mixed models are a relatively novel extension to the generalized linear mixed model, and offer the analyst tools (such as thin plate regression splines and tensor product smooths) for modeling *non-linear* functional relations between one or more predictors and a response variable. This is essential for regression modeling of a response such as the amplitude of the EEG signal, which varies nonlinearly with time.

For regression modeling—which we will need to study the effect of compound frequency as well as compound constituent frequencies—GAMMs, as implemented in the mgcv package 1.7–28, offer the possibility of modeling the EEG amplitude as a nonlinear function of time and frequency simultaneously, resulting in potentially wiggly surfaces (or, in case of more than two numerical predictors, in wiggly hypersurfaces). By decomposing the EEG amplitude into a sequence of additive components, GAMMs afford the analyst a toolkit for separating out (potentially non-linear) partial effects due to different kinds of predictors (e.g., language group, time, compound frequency, constituent frequency).

[7]**Table A6** in the Appendix gives a summary of the random effects for this model.

[8]The grand average ERPs for the raw data can be found in **Figure A1** in the Appendix.

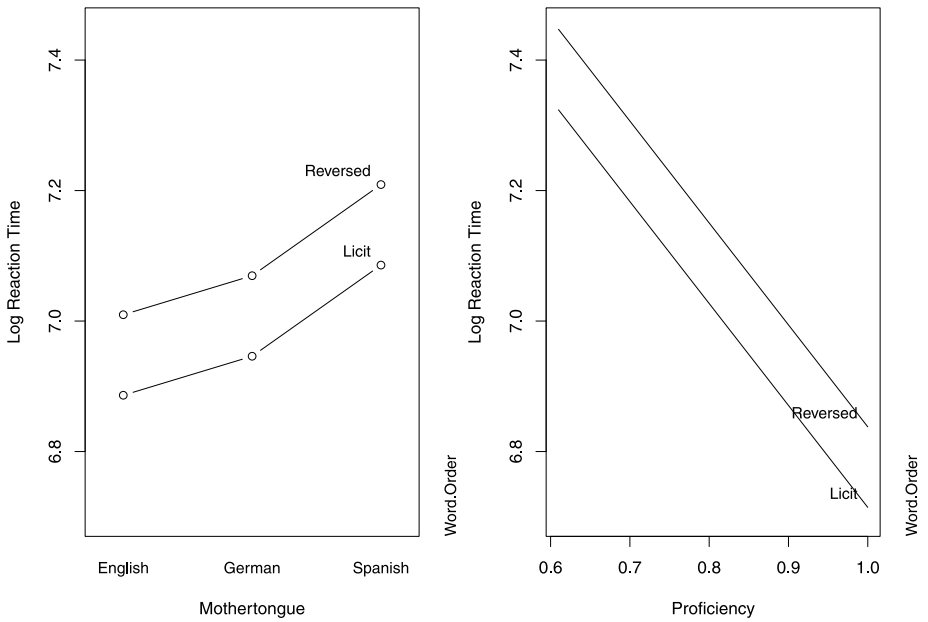


FIGURE 2 | Partial effects of the predictors in the mixed-effects model fit to the log-transformed response latencies in Study 1. The left panel is calibrated for the reference levels of Word Order (Licit) and L1 (English), and the right panel is calibrated for median proficiency.

Table 5 | Coefficients of a logistic mixed-effects regression model fitted to the accuracy data.

	Coefficient	Std. Error	Z	p
Intercept	−0.8676	1.7482	−0.4963	0.6197
Word.Order:Reversed	0.3013	0.4174	0.7218	0.4704
L1: German	0.2444	0.4981	0.4907	0.6236
L1: Spanish	0.0142	0.3072	0.0462	0.9631
Proficiency	3.8813	1.7261	2.2487	0.0245
Word.Order:Reversed by L1: German	−0.9381	0.4039	−2.3224	0.0202
Word.Order:Reversed by L1: Spanish	−0.9426	0.3999	−2.3571	0.0184

The reference level for Word.Order is Licit, and for L1: English.

In addition, GAMMs can capture AR1 autocorrelative processes in the signal, and therefore offer some protection against anti-conservative *p*-values and mistakingly taking noise for complex ERP signatures (as has been shown to occur by Tanner et al., 2013)[9]. For the present analysis, most autocorrelative structure in the residual error was removed by including in the GAMM an autocorrelation parameter $\rho = 0.9$ for AR1 error for each basic time series in the data (the time series amplitudes for each unique combination of subject and item). Thanks to inclusion of the ρ parameter, there was little remaining autocorrelation in the model's residuals, as required.

[9]Using *post-hoc* correlation analyses, (Tanner et al., 2013) found that grand mean waveforms showing a biphasic N400 + P600 response in fact concealed a more complex pattern, in which most individuals showed either an N400 or a P600, but not both.

Finally, we analyzed the EEG amplitude without any prior aggregation, seeking to predict the development of the EEG amplitude over time for any individual combination of subject and item. With 609,500 observations at each channel, we refrained from fitting a single GAMM to the full dataset. Instead, we fitted a separate GAMM to individual channels (i.e., the electrodes were analyzed independently), expecting to find similar regression curves and regression surfaces at neighboring channels. In other words, precisely because channels are not independent, topographical consistency can be relied upon as a criterion for having confidence in the regression effects.

The GAMMs provided by the `mgcv` package are designed to work fluently with treatment coding for factorial predictors. In order to inspect potential interactions between L1 group (three levels) and Word Order (two levels), we created a new six-level factor, which we labeled OG ("ordered grouping"), with levels English:Licit, English:Reversed, German:Licit, German:Reversed, Spanish:Licit, and Spanish:Reversed, with English:Licit as reference level.

Thus, we modeled the amplitude of the EEG signal (without any prior averaging) as an additive function of the fixed-effect factor OG and three covariates: Compound Frequency, and the Constituent Frequencies of Modifier and Head. Proficiency did not reach significance and did not improve the model fit significantly, so we did not include this covariate in the final model.

Participant and Compound were included in the model as random-effect factors. For Compound, we included random intercepts, in order to allow for differences in baseline amplitude across compounds. For Participant, we included two separate random-effects structures: a nonlinear factor smooth for Trial, and a second nonlinear factor smooth for Time. (These factor smooths are the non-linear counterpart of what in a strictly

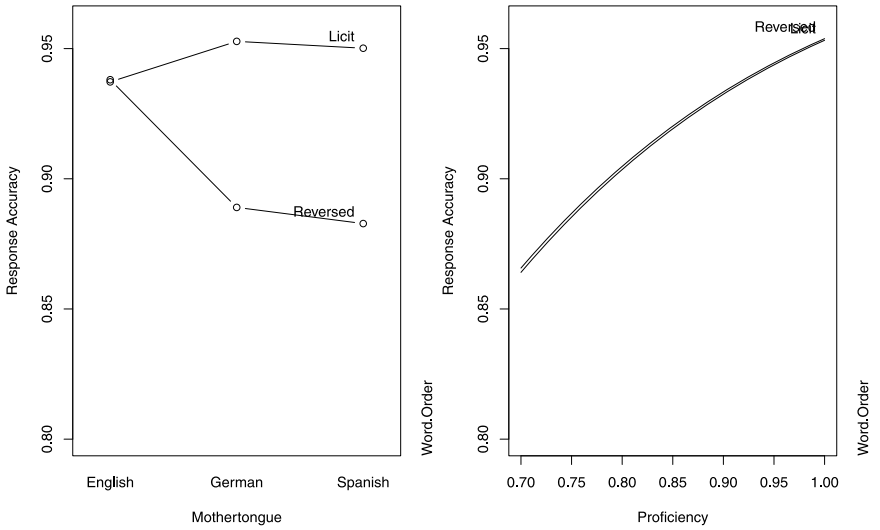


FIGURE 3 | Partial effects of the predictors in the logistic model for response accuracy in Study 2 (delayed primed lexical decision). The left panel is calibrated for the reference levels of Word Order (Licit) and L1 (English), and the right panel is calibrated for median proficiency.

linear model would have to be modeled by the combination of random intercepts and random slopes, i.e., by-participant calibration of regression lines.) The factor smooths for Trial model the development of a subject's amplitude over the course of the experiment. The factor smooths for Time model a subject's typical development of the EEG amplitude while being exposed to a given compound. These factor smooths typically afford substantial improvement to the model fit, but as these smooths are not of theoretical interest in the framework of this study, we do not discuss them in detail.

Table 6 presents a summary of the GAMM fitted to the EEG amplitude at channel C3[10]. The upper half of this table presents the parametric part of the model, with coefficients familiar from standard linear modeling with treatment coding for factors. The first six rows present the intercept (representing the group mean for English speakers in the licit word order condition, for log-transformed compound and constituent frequencies equal to 0), and the changes in the intercept for the five other factor levels. The only significant difference pertains to English speakers responding to compounds with reversed word order. In this condition, the mean amplitude was shifted down by 0.64. The second six rows summarize the effect of (log) Compound Frequency, which turned out to be linear. For English speakers presented with compounds with normal constituent order, a greater compound frequency predicted lower-valued amplitudes. The differences in slope for the other five combinations of group and word order indicate that here the slopes for Compound Frequency were around zero. For instance, for the English Reversed condition, the slope was $-0.14 + 0.17 = 0.03$. A separate model (not shown) testing the six slopes against zero revealed, as expected, a significant negative slope for licit compounds in English, and also a reduced negative slope (-0.078) for reversed compound for Spanish speakers ($p = 0.0414$). Thus, the Spanish speakers show, for the reversed condition, a pattern that resembles, albeit in weakened form, the pattern observed for English in the licit condition. Recall that in the non-delayed lexical decision task (Study 1), Spanish speakers responded with reduced accuracy in the reversed condition, compared to English speakers. Since in Spanish, the reversed word order would be the licit order, we may be seeing in the EEG amplitude the consequences of expecting (given one's L1 experience) a given constituent order (the licit order for English, but the reversed order for Spanish speakers).

The second half of **Table 6** describes the thin plate regression spline smooths (first six rows) for the development of the amplitude over time, the nonlinear interaction of the compound's constituent frequency (second six rows), and the random-effect structure in the model (last three rows)[11]. The column labeled `edf` presents the *effective degrees of freedom*: smooths with higher edf tend to be more wiggly. The first smooth, for English in the licit condition, presents the development of the amplitude over time for the corresponding subset of the data. The next 5 rows evaluate difference curves with respect to the English licit condition. The summary indicates that there are significant differences between English licit and the other combinations of Group and Word Order, with the exception of Spanish in the reversed Word

[10] 31 other models were fitted, one per chanel. Each single-chanel analysis was carried out on 609,500 data points. The main results of these models are summarized by means of **Figures 4**, **5**. Patterns that show geographical consistency across neighboring channels are the ones we have most confidence in. We focus on C3 in the model presentation, as a representative chanel for the effects of interest in our study. A baseline period was not included in the figures, because the pre-target window is one for which differential effects are expected, as different primes are presented. At −100 ms before the target word, the prime is still being read (−100 to −50 ms, followed by 50 ms of mask). Baselining has been carried out to nullify intercept shifts due to the prime, but we do NOT expect the same profile across conditions, because the primes are different, and related to the compounds in different ways.

[11] The random effect for participant over time is plotted in **Figure A2** in the Appendix.

Table 6 | Generalized additive mixed model fitted to the amplitude of the electrophysiological response of the brain to English compounds at channel C3.

A. Parametric coefficients	**Estimate**	**Std. Error**	***t*-value**	***p*-value**
Intercept (English licit)	0.5974	0.6793	0.8794	0.3792
Intercept Δ English reversed	−0.6369	0.1657	−3.8440	0.0001
Intercept Δ German licit	−1.2366	0.9314	−1.3276	0.1843
Intercept Δ German reversed	−1.5237	0.9320	−1.6348	0.1021
Intercept Δ Spanish licit	−0.2747	0.9336	−0.2942	0.7686
Intercept Δ Spanish reversed	0.6333	0.9322	0.6794	0.4969
Compound frequency (English licit)	−0.1385	0.0368	−3.7636	0.0002
Compound frequency: Δ English reversed	0.1731	0.0350	4.9499	<0.0001
Compound frequency: Δ German licit	0.1117	0.0352	3.1748	0.0015
Compound frequency: Δ German reversed	0.1154	0.0359	3.2122	0.0013
Compound frequency: Δ Spanish licit	0.1971	0.0354	5.5691	<0.0001
Compound frequency: Δ Spanish reversed	0.0606	0.0363	1.6685	0.0952
B. Smooth terms	**edf**	**Ref.df**	***F*-value**	***p*-value**
Spline smooth time (English licit)	8.5375	8.6981	12.3205	<0.0001
Spline smooth time: Δ English reversed	3.3899	4.3034	6.5872	<0.0001
Spline smooth time: Δ German licit	1.0013	1.0018	3.8845	0.0487
Spline smooth time: Δ German reversed	4.1062	5.1882	3.1005	0.0078
Spline smooth time: Δ Spanish licit	3.9976	5.0409	6.8602	<0.0001
Spline smooth time: Δ Spanish reversed	1.0227	1.0320	0.9527	0.3293
Tensor smooth freq C1, Freq C2 (English licit)	9.9401	10.6705	4.1504	<0.0001
Tensor smooth freq C1, Freq C2: Δ English:Reversed	7.4023	8.5028	4.6581	<0.0001
Tensor smooth freq C1, Freq C2: Δ German:Licit	11.7144	12.3939	8.8861	<0.0001
Tensor smooth freq C1, Freq C2: Δ German:Reversed	6.9721	8.1846	4.9458	<0.0001
Tensor smooth freq C1, Freq C2: Δ Spanish:Licit	9.4385	10.4868	11.6824	<0.0001
Tensor smooth freq C1, Freq C2: Δ Spanish:Reversed	9.5047	10.6210	4.3967	<0.0001
Smooth item (Compound)	93.0669	111.0000	6.6982	<0.0001
Smooth trial by participant	141.1186	267.0000	8.1540	<0.0001
Smooth time by participant	186.7179	266.0000	4.4254	<0.0001

Treatment coding was used for the six-level factor for the interaction of L1 by Word Order, with English Licit as reference level.

Order. As observed above for Compound Frequency, the Spanish in the reversed condition again pattern with the English in the licit condition.

The nonlinear interaction of the constituent frequencies by OG was modeled analogously, with a tensor smooth for English Licit, and difference smooths for the other levels of OG. As can be read of **Table 6**, all difference smooths reached significance.

To understand what the spline and tensor smooths represent, visualization is essential. Although visualization of the present model is straightforward, it pitches the Spanish and German, as well as the English reversed condition against the English Licit condition. Given that we have established the presence of many significant differences with English compounds in their normal word order as read by native speakers of English, we proceed with visualization based on the same model but fitted to the individual languages, contrasting the licit condition with the reversed condition (the output models are not presented in the text nor tables).

Figures 4, **5** present a summary overview of the regression curves and surfaces obtained. Within each plot region (upper rows: English, middle rows: German, bottom rows: Spanish; left column: the licit condition; right column: the difference curve (or surface) for the reversed condition).

Within a plot region, panels are arranged roughly following the topography of the EEG cap, with frontal channels at the top and parietal channels at the bottom. Only those channels are shown for which the effect was significant ($p < 0.01$).

First consider the right-hand half of **Figure 4**, focusing on the violation condition (in which compound constituents were presented in reversed order). The upper panel of plots shows a negative inflection in the difference curve around 200–400 ms post stimulus onset at left frontal and central channels for the English speakers. A more pronounced negative inflection starting around 400 ms post stimulus onset is visible for the German speakers (center left panel), again at left frontal and central sites. Interestingly, at right frontal sites, this negative inflection reverses into a strong positivity. For the Spanish speakers, left frontal and midline channels show a reduced but still significant negative inflection in the difference curve, also starting around 400 ms. This suggests an early N400 effect for English speakers, and a standard N400 effect for the non-native groups (although delayed,

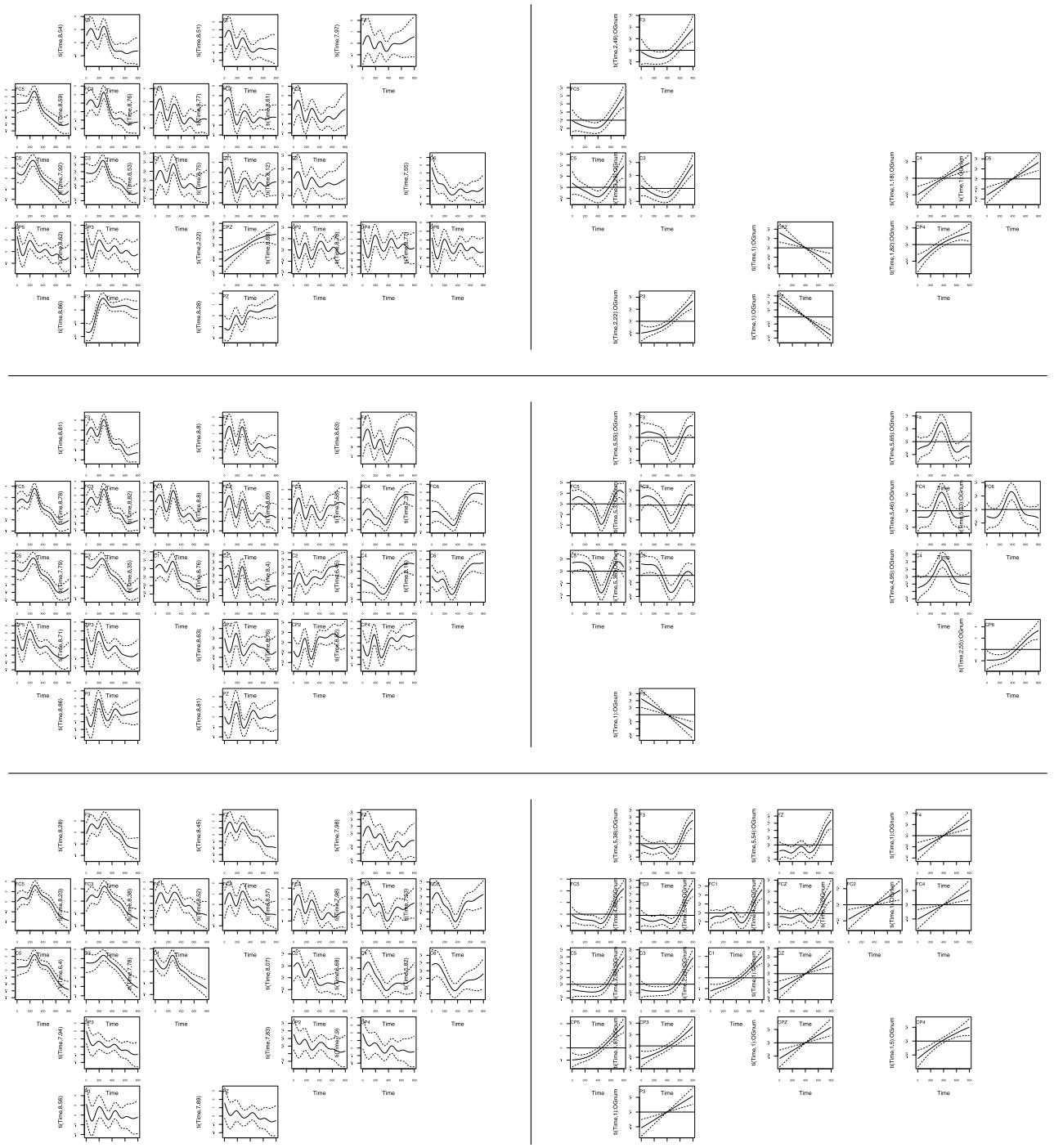


FIGURE 4 | The three-way interaction of Participant Group, Word Order, and Time. Upper row: English, middle row: German, bottom row: Spanish; left column: amplitude development over time for the **licit** constituent order, right column: the difference curve for the **reversed** word order, with confidence intervals (dotted lines). Details of individual panels can be inspected by zooming in with higher magnification.

as expected for non-natives—Moreno and Kutas 2005), with the strongest effect emerging for the German speakers[12].

The N400 is traditionally considered to reflect semantic integration processes (Kutas and Federmeier, 2011), and its amplitude has been found to be larger for non-words than words (Kutas and Federmeier, 2000), including when the test items were (reversed and non-reversed) compounds (El Yagoubi et al.,

[12]But see Discussion for an alternative explanation of the observed negative inflection as LAN.

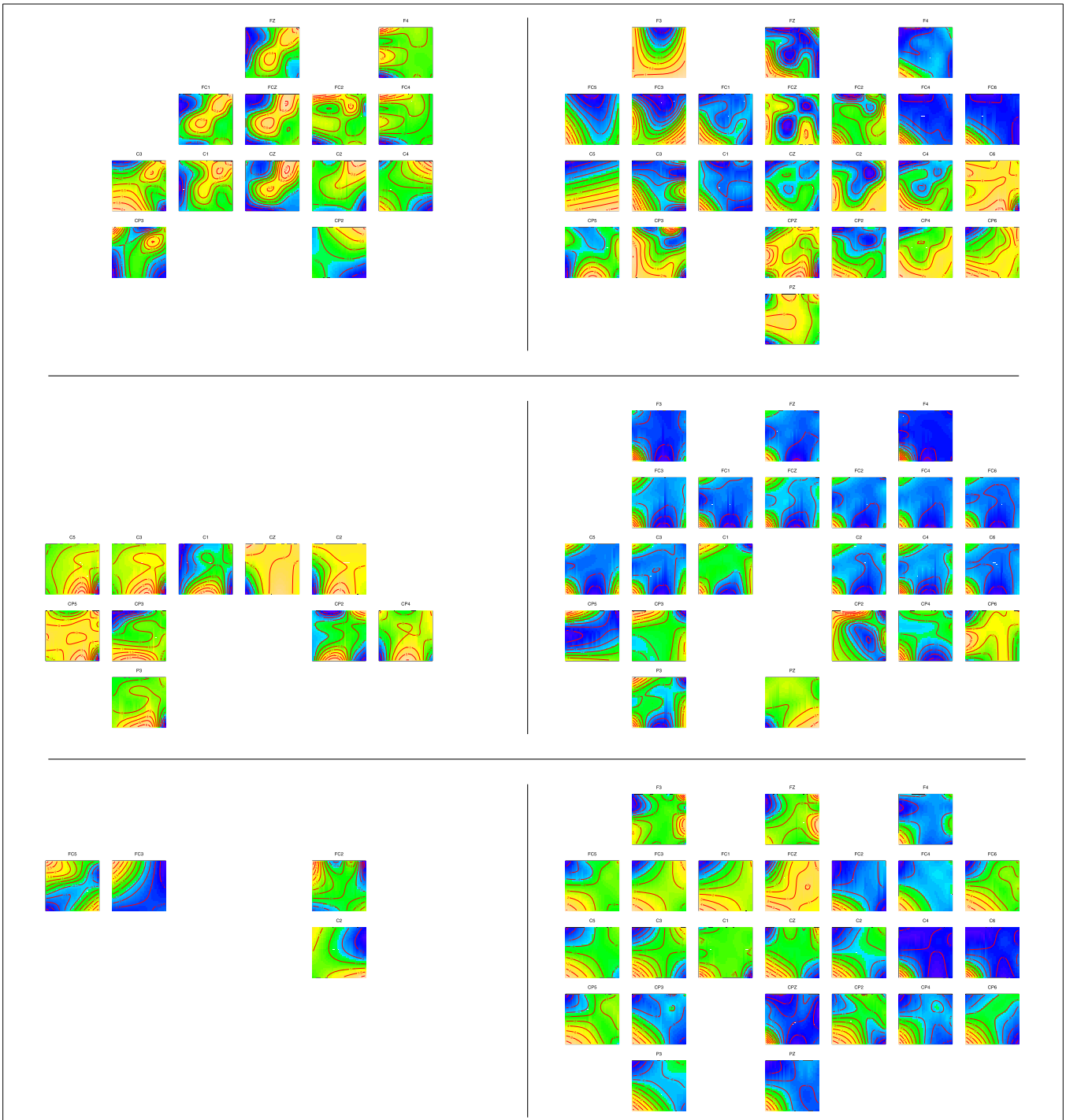


FIGURE 5 | The four-way interaction of Participant Group, Word Order, and the two constituent frequencies (horizontal: frequency of first noun, vertical: frequency of second noun). Upper row: English, middle row: German, bottom row: Spanish; left column: amplitude surface for the licit constituent order, right column: the difference surface for the reversed word order. Channels where there was no significant effect are not shown. Top panels present frontal channels, whereas the bottom panels show the parietal channels. Darker shades of blue indicate larger negative partial effects, whereas yellow and white denote larger positive partial effects. Details of individual panels can be inspected by zooming in with higher magnification.

2008). This ERP signature traditionally reported at more parietal electrodes, but (Voss and Federmeier, 2011) demonstrated that it can also be found in more anterior locations, as we do here.

All groups featured a significant positive peak in amplitude around 300 ms, as can be seen in the left plot regions of **Figure 4**. As the difference curves in the corresponding right plot regions are relatively flat for the first 300 ms, this P300 also characterized

the reading of compounds with reversed word order. This effect was more pronounced for English and German speakers, and somewhat attenuated for the Spanish speakers. In all groups, this peak occurred earlier at more parietal regions in the left hemisphere, suggesting a possible spreading from parietal to frontal regions.

In the Reversed Word Order condition, the English and Spanish groups feature a significant positive inflection in amplitude at left frontal sites starting around 500 ms and rising up to the end of the time window [0-800 ms], suggesting a higher, later peak. The German group does not feature this robust pattern. Furthermore, the English and Spanish, but not the Germans, show at some right channels a linear increase in amplitude over time.

Summing up, the violation of English word order is reflected in the EEG signal by an N400 effect. For English and Spanish, a positivity around 600 ms post stimulus onset may reflect a P600 (or perhaps a P500) indexing the processing of syntactic repair or integration (Kaan, 2007). Comparing the three groups, the Spanish difference curves group together with the English difference curves, whereas the German difference curves stand apart with a stronger N400 effect at left frontal sites and, surprisingly, a P400 effect at right frontal sites.

Figure 5 presents the three-way interaction of the frequency of the first constituent (horizontal axis of each contour plot) by the frequency of the second constituent (vertical axis of each contour plot) by OG. Darker shades of blue indicate larger negative partial effects, whereas yellow and white denote larger positive partial effects.

First consider channel C3 in the upper left panel of plots of **Figure 5**. What this panel shows is that higher amplitudes are characteristic for compounds for which both constituent frequencies are either high (upper right corner) or low (lower left corner). Lower amplitudes are characteristic for mismatching constituent frequencies. This kind of cross-over interaction has been observed previously for the constituents of derived words in an eye-tracking study of reading (Kuperman et al., 2010), suggesting that an imbalance in constituent frequencies increases entropy, leading to an increased processing load.

This cross-over effect, which is also visible at neighboring channels (FZ, FCZ, FC1, C1, Cz, C2, C4) is present only for English readers in the licit condition. German speakers in the licit condition (center left panel) show an inverse U-shaped effect of modifier frequency for lower values of head frequency at most channels. We think this effect may be the result of the prior priming of the head constituent, which may have affected the nonnative speakers of German more than the native speakers of English. The inverse U-shaped effect may represent optimization of the response to those words which have probabilities (gauged by their corpus frequencies) that are themselves probable, i.e., in the center of the (lognormal) probability distribution. In other words, we think it is not the relative frequency of the modifier itself that predicts the amplitude, but the probability of that relative frequency.

For Spanish, significant results for the licit word order (shown in the lower left plot region) are too scattered to provide a realistic basis for interpretation.

Next consider the consequences of reversing constituent order, as shown in the right-hand half of **Figure 5**. For English and German (top and center panels), and more right-lateralized for Spanish (lower panel), downward adjustments of the amplitude are widespread, especially at more frontal sites in the English and German groups. We speculate that source analysis will find that these negativities reflect conflict resolution processes originating from the anterior cingulate cortex (ACC) Botvinick et al. (2001); Yeung et al. (2004): the constituents are legitimate, but their order is not, resulting in conflicting evidence for a lexicality decision. Note that the kind of "conflicts" that arise due to what is generally described as lexical competition (e.g., neighbors) is qualitatively different from the conflict arising with our experimental manipulation, which involves higher-order meaningful constituents that in half of the trials are saliently out of order.

For English, patterns across channels vary widely, with the common feature that negative effects are pervasive for high head frequencies. Since the head was primed, the appearance of the head in the inconventional initial position may have induced greater processing costs especially for higher-frequency heads.

The pattern for German (center right plot region) is much more systematic. The inverse U-shaped effect that emerged for the licit word order is negated by a U-shaped negative inflection of the EEG wave. This negative inflection is even present at many sites where no significant effect was discernable in the licit condition (see e.g., all F and FC channels). The change in polarity of the effect suggests the hypothesis that the negative, downwards, adjustments to the EEG waveform are an index of processing costs, whereas the positive (inverse U-shaped) effects in the licit condition reflect facilitated processing.

The pattern for Spanish in the reversed condition is strikingly different from that for English and German. First, the sensors in the left hemisphere reveal a pattern that bears some resemblance to the pattern for English in the licit condition, compare for instance C3 for English licit and Spanish Reversed. Compounds with constituents of similar frequency show positive inflections, whereas constituents of dissimilar frequency show negative inflections. Since the negative inflections correspond to high-entropy situations, this pattern fits nicely with the hypothesis advanced above that positive inflections reflect facilitated processing, and negative inflections, increased processing costs. The reason that the Spanish in the reversed condition pattern with the English in the licit condition is most likely to be the licitness of the reversed word order for Spanish.

Interestingly, the negative effects at many channels in the right hemisphere, as well as at more parietal channels, set the Spanish apart from English in both the licit and reversed word order conditions. We think these negativities reflect the processing invested in resolving the incongruity of the licit Spanish word order for English compounds.

5. DISCUSSION

The present examination of similarities and differences between native and non-native reading of English compounds revealed

Table 7 | Summary of Results.

	English	Spanish	German
Speeded RT (Study 1)	Short	Long	Short
Speeded accuracy (Study 1)	High	Low (Reversed)	High
Appearance N400	Early	Late	Late
Presence P500/P600	Yes	No	Yes
Compound frequency	Yes (Licit)	Yes (Reversed)	No
Crossover effect of Constituent frequencies	Yes (Licit)	Yes (Reversed)	No
U-shaped modifier Frequency effect	No	No	Yes (Polarity with word order)
Pervasive frontal Negativity	Yes	No (only right hemisphere)	Yes

the results summarized in **Table 7**.[13]First, in the speeded lexical decision task (Study 1), the L2 participants' accuracy rates for compounds with licit constituent order were indistinguishable from those of native speakers of English. This indicates that the target structure of English compounds has been acquired, and that there is no representational deficit. This is unsurprising as no functional morphology is involved (Lardiere, 2008; Slabakova, 2008) and head-directionality transfer effects are expected to be short-lived (Haznedar, 1997; Unsworth, 2005).

For compounds presented with reversed constituent order, performance dropped for all groups (except the native group in the delayed lexical decision task—Study 2). Typically, errors consisted of the over-acceptance of reversed compounds, and would be classified as 'false alarms' rather than "misses" in Detection Theory (Macmillan and Creelman, 2005). Whereas accuracy of German speakers was very similar to that of English speakers, the accuracy of Spanish speakers was significantly reduced under word order reversal. Furthermore, it was only for the Spanish speakers that response latencies were significantly slower than those of English native speakers (in Study 1), a result not expected according to the Interface Hypothesis (Sorace, 2011)—which predicts similar processing difficulties in the non-native language, irrespective of the properties of the L1. The slower responses of the Spanish L2 speakers suggest an interference effect from their native language: Rejecting a compound presented in reversed order requires the Spanish participants to reject what would be a licit word order in their L1. This is where they make errors, and where their responses become elongated. These results reveal the presence of L1-induced residual errors in the processing of a core grammar phenomenon.

The ERP results for Spanish fit well the presence of an L1 effect. The Spanish speakers show an effect of compound frequency, just as the English speakers, but for the reversed (i.e., their native) word order. The Spanish speakers also show a crossover effect of the constituent frequencies, as do the English speakers, again for the reversed instead of the licit word order. The compound frequency effect suggests familiarity with the onomasiological function of the compound when the constituents appear in the order appropriate for their L1. The crossover effect of the constituent frequencies is likewise conditioned on the order in the speakers' L1, and may bear witness to higher processing costs when the entropy of the probability distribution of modifier and head [as gauged by their (relative) frequencies] is high (see Kuperman et al., 2010).

The frequency effects present for the German speakers are very different from both those of English and of Spanish speakers. Their EEG signal was not predictable from compound frequency, suggesting decomposition (i.e., full parsing). Furthermore, the constituent frequency effects were different in nature, showing for modifier frequency (conditional on a low head frequency) an inverse U-shaped curve for licit word order, and a U-shaped pattern for the reversed word order. For these speakers, the violation condition is characterized by topographically pervasive negativities. This suggests that German speakers were especially sensitive to the word order violation in English, which also violates the expected word order in German. Support for this hightened sensitivity comes from the N400 effect for this group of speakers, which is characterized by a well-defined narrow large downward inflection for the reversed compounds. Of course, the speakers of the other two languages must also have been aware of the violations, as indicated by their increased error rates and longer response latencies. Nevertheless, the N400 effects for the English and Spanish speakers are not as pronounced as for the German speakers. A final difference between the German speakers and the other two language groups, for which we have no explanation, is the absence of a clear positivity starting around 600 ms post stimulus onset (possibly a P500 or a P600 effect indexing reanalysis and repair), and the presence of a positive inflection around 400 ms post stimulus onset at channels at right frontal sites, the mirror image of the N400 effect.

An alternative interpretation for the negativity observed around 400 ms post-stimulus onset in the present study is that it reflects the left anterior negativity (LAN) component which is assumed to index integration of morphosyntactic information (Friederici, 1995, 2001; Steinhauer et al., 2009)[14]. In fact, the scalp distribution of the observed component (anterior and predominantly left) does align with LAN. The LAN has been shown to be elicited by subject verb agreement violations (but not by number or gender violations between an antecedent and a reflexive pronoun—Osterhout and Mobley 1995), grammatical gender violation (Gunter et al., 2000), and pronoun case and verb agreement errors (Coulson et al., 1998). Though the LAN component is typically observed in studies with sentence stimuli, it is possible to interpret our findings as a LAN if we assume that the processing and violations in the compounds used in the present study are morpho-syntactic rather than semantic in nature. Assuming that the anterior negativity is LAN, rather than N400, and indexes morpho-syntactic processing rather than semantic processing, the results are consistent with

[13]With a high value for the ρ parameter, our analyses are conservative. Furthermore, with a Bonferroni correction for 32 channels by 27 coefficients or smooth terms, any term in **Table 6** for which $p < 0.0001$ is reasonably well supported. Nevertheless, with only 10 speakers for each group, only a replication study can reveal how robust the regression curves and regression surfaces actually are.

[14]Thanks to an anonymous reviewer for suggesting this.

El Yagoubi et al. (2008), who found a more negative peak in the left anterior negativity (LAN) component for compounds than for noncompounds. Arcara et al. (2014) further reported an enhanced LAN in head-final compounds in Italian, which they argue indicates they are decomposed differently to head-initial compounds (the latter being seemingly processed as syntactic-like structures rather than morphological complex words). LAN modulation has also been noted in two ERP papers on German compound processing (Koester et al., 2004, 2007). These researchers argued for compound decomposition during comprehension providing evidence against full-listing models and in favor of decomposition or dual-route models of compound processing.

The P300 effect that we observed for all participant groups in both word order conditions could be linked to the binary decision (licit/illicit) the participants had to make regarding the stimuli (Donchin and Coles, 1988; Barber and Carreiras, 2005). Thus, regardless of whether the stimuli were licit or illicit, participants had to attend and indicate their decision: the P300 here could be interpreted as indexing attention associated with language processing. Several authors have proposed that P300 activity is related to subsequent P600 activity for reanalysis and repair processes (e.g., Friederici, 1995).

All groups were sensitive to the probabilities of the modifier and head constituents. This challenges the claim of Silva-Corvalan and Clahsen (2008) that non-native speakers would rely on whole-word processing without understanding the constituents, but is consistent with a syntactic analysis of noun-noun compounds. Our results suggest that lexically transparent NNCs with low frequencies are processed combinatorially by (advanced) non-native speakers, as they are by native speakers (MacGregor and Shtyrov, 2013). Our findings are also consistent with the conjoint effects of both whole-word and constituent probabilities in the eye-tracking record, as early as first fixation durations (see, e.g., Kuperman et al., 2008, 2009; Miwa et al., 2014, for English, Finnish, and Japanese respectively). The importance of the constituents for non-native speakers is reminiscent of the decompositional eye-movement patterns of less-proficient readers reported by (Kuperman and Van Dyke, 2011).

Our study confirms the importance of the Third Factor (Chomsky, 2005) in L2 research: it suggests that processing effects can be induced by properties of the L1 that cannot be fully inhibited during L2 processing, in spite of acquisition of the target representation. In terms of Detection Theory (Macmillan and Creelman, 2005), this predicts that false alarms (i.e., accepting an illicit structure) will persist when misses (i.e., failing to accept a licit structure) have dropped to non-significant levels. It might be that domain-general inhibition is required to suppress L1 interferences in L2 processing, in the same way as it is recruited for language switching (de Bruin et al., 2014), in which case a corrleation would be expected between the rate of false alarms and inhibition abilities (all other things being equal).

Methodologically, the insights gleaned from the EEG amplitudes would not have been possible without generalized additive mixed models. At the same time, we believe we are only seeing the tip of the iceberg. For instance, the model can be improved by allowing the interaction of the constituent frequencies by group and constituent order to vary with time, using five-way tensor product smooths. Two considerations have withheld us from following up on such considerably more complex models. First, without specific hypotheses as a guide, interpretation becomes extremely difficult. Second, we are concerned that with a relative small number of compounds (120), overfitting might become an issue. For future research specifically addressing the development over time of constituent (and whole-compound) frequency effects, we recommend regression designs with substantially larger numbers of compounds. Replication studies will be essential for boosting confidence in the nonlinear effects revealed by the GAMMs.

AUTHOR CONTRIBUTIONS

Cecile De Cat: The first author conceived the project and was substantially involved in all aspects of its design and realization (except for data collection), as well as in the analysis and interpretation, and the drafting and revision of the manuscript. Ekaterini Klepousniotou: The second author contributed substantially to the design and realization, oversaw the data collection and initial data preparation, contributed to the interpretation of the results and critically revised the manuscript. R. Harald Baayen: The third author led and substantially contributed to the analysis of the ERP data and its interpretation, and contributed substantially to the drafting of the relevant sections and conclusions. All authors are responsible for final approval of the version to be published and agree to be accountable for all the aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

The third author was supported by an Alexander von Humboldt research chair awarded by the Alexander von Humboldt foundation, and the first author was supported by a British Academy Skills Acquisition award (SQ120066) and by the Leeds Humanities Research Institute.

ACKNOWLEDGMENTS

Many thanks to Antoine Tremblay for his generous help with the data preparation script, to Cyrus Shaoul for friendly technical and coding advice, to Jacolien van Rij for helpful suggestions for the GAMM analysis, to the anonymous reviewers for their insightful comments, and to Raphael Morschett, Chris Norton, Kremena Koleva and Natasha Rust for the data collection and pre-processing.

REFERENCES

Arcara, G., Marelli, M., Buodo, G., and Mondini, S. (2014). Compound headedness in the mental lexicon: an event-related potential study. *Cogn. Neuropsychol.* 31, 164–183. doi: 10.1080/02643294.2013.847076

Barber, H., and Carreiras, M. (2005). Grammatical gender and number agreement in spanish: an erp comparison. *J. Cogn. Neurosci.* 17. 137–153. doi: 10.1162/0898929052880101

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. R package version 1.0-4.

Bertram, R., Hyönä, J., and Pollatsek, A. (2004). Morphological parsing and the use of segmentation cues in reading Finnish compounds. *J. Mem. Lang.* 51, 325–345. doi: 10.1016/j.jml.2004.06.005

Botvinick, M., Braver, T., Carter, C., Barch, D., and Cohen, J. (2001). Evaluating the demand for control: anterior cingulate cortex and crosstalk monitoring. *Psychol. Rev.* 108, 624–652. doi: 10.1037/0033-295X.108.3.624

Briggs, G., and Nebes, R. (1975). Patterns of hand preference in a student population. *Cortex* 11, 230–238. doi: 10.1016/S0010-9452(75)80005-0

Butterworth, B. (1983). "Lexical representation," in *Language Production*, ed B. Butterworth (San Diego, CA: Academic Press), 257–294.

Chomsky, N. (2005). Three factors in language design. *Linguist. Inq.* 36, 1–22. doi: 10.1162/0024389052993655

Clahsen, H., Balkhair, L., Schutter, J.-S., and Cunnings, I. (2013). The time course of morphological processing in a second language. *Second Lang. Res.* 29, 7–31. doi: 10.1177/0267658312464970

Coulson, S., King, J., and Kutas, M. (1998). Expect the unexpected: event-related brain response to morphosyntactic violations. *Lang. Cogn. Process.* 13, 21–58. doi: 10.1080/016909698386582

de Bruin, A., Roelofs, A., Dijkstra, T., and FitzPatrick, I. (2014). Domain-general inhibition areas of the brain are involved in language switching: fmri evidence from trilingual speakers. *NeuroImage* 90, 348–359. doi: 10.1016/j.neuroimage.2013.12.049

Donchin, E., and Coles, M. (1988). Is the p300 component a manifestation of context updating? *Behav. Brain Sci.* 11, 357–374. doi: 10.1017/S0140525X00058027

El Yagoubi, R., Chiarelli, V., Mondini, S., Perrone, G., Danieli, M., and Semenza, C. (2008). Neural correlates of Italian nominal compounds and potential impacts of headedness effect: an ERP study. *Cogn. Neuropsychol.* 25, 559–581. doi: 10.1080/02643290801900941

Friederici, A. D. (1995). The time course of syntactic activation during language processing: a model based on neuropsychological and neurophysiological data. *Brain Lang.* 50, 259–281. doi: 10.1006/brln.1995.1048

Friederici, A. D. (2001). "Event-related brain potentials and aphasia," in *Handbook of Neuropsychology*, 2nd Edn., Vol. 3, eds F. Boller and J. Grafman (Amsterdam: Elsevier Science), 353–373.

Gagné, C. L., and Spalding, T. L. (2014). Conceptual composition: the role of relational competition in the comprehension of modifier-noun phrases and noun-noun compounds. *Psychol. Learn. Motiv.* 59, 97–130. doi: 10.1016/B978-0-12-407187-2.00003-4

Gunter, T., Friederici, A., and Schriefers, H. (2000). Syntactic gender and semantic expectancy: erps reveal early autonomy and late interaction. *J. Cogn. Neurosci.* 12, 556–568. doi: 10.1162/089892900562336

Haznedar, B. (1997). *Child Second Language Acquisition of English: A longitudinal Case Study of a Turkish-Speaking Child*. Doctoral dissertation, Durham University, Durham.

Hyönä, J., and Pollatsek, A. (1998). Reading finnish compound words: eye fixations are affected by component morphemes. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1612–1627. doi: 10.1037/0096-1523.24.6.1612

Jarema, G. (2006). "Compound representation and processing: a cross-language perspective," in *The Representation and Processing of Compound Words*, eds G. Libben and G. Jarema (Oxford: OUP), 45–70.

Jarema, G., Busson, C., Nikolova, R., Tsapkini, K., and Libben, G. (1999). Processing compounds: a cross-linguistic study. *Brain Lang.* 68, 362–369. doi: 10.1006/brln.1999.2088

Juhasz, B., Starr, M., Inhoff, A., and Placke, L. (2003). The effects of morphology on the processing of compound words: evidence from lexical decision, naming, and eye fixations. *Br. J. Psychol.* 94, 223–244. doi: 10.1348/000712603321661903

Kaan, E. (2007). Event-related potentials and language processing. a brief introduction. *Lang. Linguist. Compass* 1, 571–591. doi: 10.1111/j.1749-818X.2007.00037.x

Koester, D., Gunter, T. C., and Wagner, S. (2007). The morphosyntactic decomposition and semantic composition of german compound words investigated by erps. *Brain Lang.* 102, 64–79. doi: 10.1016/j.bandl.2006.09.003

Koester, D., Gunter, T. C., Wagner, S., and Friederici, A. D. (2004). Morphosyntax, prosody, and linking elements: the auditory processing of german nominal compounds. *J. Cogn. Neurosci.* 16, 1647–1668. doi: 10.1162/0898929042568541

Kroll, J. F., Michael, E., Tokowicz, N., and Dufour, R. (2002). The development of lexical fluency in a second language. *Second Lang. Res.* 18, 137–171. doi: 10.1191/0267658302sr201oa

Krott, A., Gagne, C. L., and Nicoladis, E. (2010). Children's preference for has and located relations: a word learning bias for noun-noun compounds. *J. Child Lang.* 37, 373–394. doi: 10.1017/S0305000909009593

Kryuchkova, T., Tucker, B. V., Wurm, L., and Baayen, R. H. (2012). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Brain Lang.* 122, 81–91. doi: 10.1016/j.bandl.2012.05.005

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *Lang. Cogn. Process.* 23, 1089–1132. doi: 10.1080/01690960802193688

Kuperman, V., Bertram, R., and Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *J. Mem. Lang.* 62, 83–97. doi: 10.1016/j.jml.2009.10.001

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: towards a multiple route model of lexical processing. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 876–895. doi: 10.1037/a0013484

Kuperman, V., and Van Dyke, J. (2011). Effects of individual differences in verbal skills on eye-movement patterns duing sentence reading. *J. Mem. Lang.* 65, 42–73. doi: 10.1016/j.jml.2011.03.002

Kutas, M., and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn. Sci.* 4, 462–470. doi: 10.1016/S1364-6613(00)01560-6

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Lardiere, D. (2008). "Feature assembly in second language acquisition," in *The Role of Formal Features in Second Language Acquisition*, eds J. Liceras, H. Zobl, and H. Goodluck (New York, NY: Lawrence Erlbaum Associates), 106–140.

Libben, G. (1998). Semantic transparency in the processing of compounds. *Brain Lang.* 61, 30–44. doi: 10.1006/brln.1997.1876

Libben, G. (2006). "Why study compound processing? An overview of the issues," in *The Representation and Processing of Compound Words*, eds G. Libben and G. Jarema (Oxford: OUP), 1–22.

Libben, G., and Jarema, G., (eds.). (2006). *The Representation and Processing of Compound Words*. Oxford: OUP.

Lieber, R., and Štekauer, P., (eds.). (2009). *The Oxford Handbook of Compounding*. Oxford: Oxford University Press.

MacGregor, L. J., and Shtyrov, Y. (2013). Multiple routes for compound word processing in the brain: evidence from eeg. *Brain Lang.* 126, 217–229. doi: 10.1016/j.bandl.2013.04.002

Macmillan, N., and Creelman, C. (2005). *Detection Theory: A User's Guide*. Mahwah, NH: Lawrence Erlbaum Associates.

Marelli, M., Crepaldi, D., and Luzzatti, C. (2009). Head position and the mental representation of italian nominal compounds. *Ment. Lexicon* 4, 430–455. doi: 10.1075/ml.4.3.05mar

Marelli, M., Zonca, G., Contardi, A., and Luzzatti, C. (2014). The representation of compound headedness in the mental lexicon: a picture naming study in aphasia. *Cogn. Neuropsychol.* 31, 26–39. doi: 10.1080/02643294.2013.860024

Meyer, R. (1993). *Compound Comprehension in Isolation and in Context. The Contribution of Conceptual and Discourse Knowledge to the Comprehension of German Novel Noun-Noun Compounds*. Berlin: Walter de Gruyter.

Miwa, K., Libben, G., Dijkstra, T., and Baayen, H. (2014). The time-course of lexical activation in japanese morphographic word recognitin: evidence for a character-driven processing model. *Q. J. Exp. Psychol.* 67, 79–113. doi: 10.1080/17470218.2013.790910

Moreno, E. M., and Kutas, M. (2005). Processing semantic anomalies in two languages: an electrophysiological exploration in both languages of spanish-english bilinguals. *Brain Res. Cogn. Brain Res.* 22, 205–220. doi: 10.1016/j.cogbrainres.2004.08.010

Nicoladis, E., and Yin, H. (2002). "The role of frequency in acquisition of english and chinese compounds by bilingual children," in *Proceedings of the Annual Boston University Conference on Language Development*, Vol. 26, (Somerville, MA: Cascadilla Press), 441–452.

Osterhout, L., and Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *J. Mem. Lang.* 34, 739–773. doi: 10.1006/jmla.1995.1033

Otten, L., and Rugg, M. (2005). "Interpreting event-related brain potentials," in *Event-Related Potentials: A Methods Handbook*, ed T. Handy (Cambridge, MA: MIT Press), 3–17.

Piera, C. (1995). "On compounding in english and spanish," in *Evolution and Revolution in Linguistic Theory*, eds H. Campos and P. Kempchinsky (Washington, DC: Georgetown University Press), 301–315.

Pollatsek, A., Hyönä, J., and Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 820–833. doi: 10.1037/0096-1523.26.2.820

Sandra, D. (1990). On the representation and processing of compound words: automatic access to constituent morphemes does not occur. *Q. J. Exp. Psychol.* 42A, 529–567. doi: 10.1080/14640749008401236

Semenza, C., and Luzzatti, C. (2014). Combining words in the brain: the processing of compound words.introduction to the special issue. *Cogn. Neuropsychol.* 31, 1–7. doi: 10.1080/02643294.2014.898922

Semenza, C., Luzzatti, C., and Carabelli, S. (1997). Morphological represntation of compound nouns: a study on Italian aphasic patients. *J. Neurolinguist.* 10, 33–43. doi: 10.1016/S0911-6044(96)00019-X

Sharbrough, F., Chatrian, G., Lesser, R., Luders, H., Nuwer, M., and Picton, T. (1991). American electroencephalographic society guidelines for standard electrode position nomenclature. *J. Clin. Neurophysiol.* 8, 200–202. doi: 10.1097/00004691-199104000-00007

Silva-Corvalan, C., and Clahsen, H. (2008). Morphologically complex words in l1 and l2 processing: evidence from masked priming experiments in english. *Bilingualism* 11, 245–260. doi: 10.1017/S1366728908003404

Slabakova, R. (2008). *Meaning in the Second Language. Studies on Language Acquisition 34.* Berlin: Mouton de Gruyter.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguist. Approaches Bilingualism* 1, 1–33. doi: 10.1075/lab.1.1.01sor

Steinhauer, K., White, E. J., and Drury, J. E. (2009). Temporal dynamics of late second language acquisition: evidence from event-related brain potentials. *Second Lang. Res.* 25, 13–41. doi: 10.1177/0267658308098995

Tanner, D., Inoue, K., and Osterhout, L. (2013). Brain-based individual differences in online L2 grammatical comprehension. *Bilingualism* 17, 277–293. doi: 10.1017/S1366728913000370

Tremblay, A., and Baayen, R. H. (2010). "Holistic processing of regular four-word sequences: a behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall," in *Perspectives on Formulaic Language: Acquisition and Communication*, ed D. Wood (London: The Continuum International Publishing Group), 151–173.

Tremblay, A., and Newman, A. (2015). Modelling non-linear relationships in ERP data using mixed-effects Regression with R examples. *Psychophysiology* 52, 124–139. doi: 10.1111/psyp.12299

Unsworth, S. (2005). *Child L2, Adult L2, Child L1: Differences and Similarities. A Study on the Acquisition of Direct Object Scrambling in Dutch*. Utrecht: LOT.

Voss, J., and Federmeier, K. D. (2011). Fn400 potentials are functionally identical to n400 potentials and reflect semantic processing during recognition testing. *Psychophysiology* 48, 532–546. doi: 10.1111/j.1469-8986.2010.01085.x

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* 99, 673–686. doi: 10.1198/016214504000000980

Wood, S. (2006). *Generalised Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

Yeung, N., Botvinick, M. M., and Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol. Rev.* 111:931. doi: 10.1037/0033-295X.111.4.931

Zhang, J. I. E., Anderson, R. C., Wang, Q., Packard, J., Wu, X., Tang, S., et al. (2012). Insight into the structure of compound words among speakers of chinese and english. *Appl. Psycholinguist.* 33, 753–779. doi: 10.1017/S0142716411000555

Zipser, K. (2013). "Proto-language, phrase structure and nominal compounds. Which of them fit together?," in *Poster presented at ICL 2013* (Geneva).

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: 11 August 2014; accepted: 14 January 2015; published online: 09 February 2015.

*Citation: De Cat C, Klepousniotou E and Baayen RH (2015) Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English. Front. Psychol. **6**:77. doi: 10.3389/fpsyg.2015.00077*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

APPENDIX

Table A1 | Stimuli (in licit word order).

Adult jail	Air missile	Alcohol licence
Army depot	Ash cloud	Baby lotion
Bacon rind	Banana pancake	Bath mat
Beach ball	Bicycle bell	Bike grease
Bird virus	Blackcurrant jelly	Bread knife
Bronze manequin	Calf liver	Camp chair
Candle wick	Canine tooth	Cannabis resin
Car pollution	Cartoon series	Cattle grid
Cell nucleus	Cement block	Champagne froth
Cherry jar	Chestnut mash	Chicken leg
Church minister	Cigarette smell	Clay doll
Clothes peg	Coal dust	Coconut tree
Council leaflet	Country produce	Crime trend
Custard layer	Diary extract	Dog basket
Dress pattern	Duck poo	Ferry fume
Finance consultant	Floor tile	Flour dough
Flower petal	Flu injection	Freezer magnet
Garlic clove	Geography essay	Gold broach
Granola bar	Gravel path	Gym bag
Holiday souvenir	Home remedy	Hydrogen bubble
Ice sculpture	Ink stain	Jungle Beast
Kitchen utensil	Lace edge	Lamb kidney
Lemon zest	Lightning strike	Limestone rock
Maple leaf	Marble inkpot	Metal gate
Milk powder	Mountain goat	Music certificate
Nappy rash	Nettle juice	Nose drop
Ocean navigation	Papaya smoothie	Paper hat
Party outfit	Pen lid	Pet odour
Phone socket	Pig enclosure	Piston shaft
Plastic obstacle	Protein ingredient	Radio source
Rat poison	Rubber glove	Ruby pendant
Safety rule	Sandwich snack	Sea fish
Silver ring	Sleeve patch	Soup dish
Space debris	Sport injury	Steel rod
Stone chisel	Sun deck	Sweat band
Tea cart	Teak partition	Team mascot
Throat tablet	Timber fence	Tobacco product
Traffic noise	Travel kettle	Turtle shell
Vanilla cream	War troop	Wood preservative

Table A2 | Frequency statistics for the stimuli (in licit word order).

	Mean	SD	Median	Min	Max
Compound	359.52	640.78	96	0	3300
First constituent	279297.19	420859.13	120738	0	2759265
Second constituent	278976.08	421000.19	116003	0	2759265

Table A3 | Random effects from the logistic mixed-effects regression model fitted to the accuracy data of Study 1.

Groups	Name	Variance	Std.Dev.	Corr
Target	(Intercept)	0.7058	0.8401	
Subject	(Intercept)	0.2380	0.4879	
	Word.Order:Reversed	0.1831	0.4279	−0.08

Inclusion of by-target random slopes for group resulted in an overspecified model, and therefore was removed.

Table A4 | Random effects from the logistic mixed-effects regression model fitted to the reaction time data of Study 1.

Groups	Name	Variance	Std.Dev.	Corr	
Target	(Intercept)	0.006527	0.08079		
	L1German	0.005625	0.07500	0.12	
	L1Spanish	0.011570	0.10756	0.43	0.75
Subject	(Intercept)	0.025052	0.15828		
	Word.Order: Reversed	0.002446	0.04946	−0.28	

Table A5 | Mean Reaction Times (in ms) by Participant Group and Word Order condition in the speeded lexical decision task (Study 1).

	English	German	Spanish
Licit word order	877.07	1473.11	1566.14
Reversed word order	973.99	1588.83	1625.23

Table A6 | Random effects from the logistic mixed-effects regression model fitted to the reaction time data of Study 2.

Groups	Name	Variance	Std.Dev.	Corr	
Target	(Intercept)	0.9445	0.9718		
	L1German	0.7516	0.8669	−0.44	
	L1Spanish	0.7443	0.8627	−0.29	0.75
Participant	(Intercept)	0.5815	0.7626		
	Word.Order Reversed	0.6033	0.7767	−0.71	

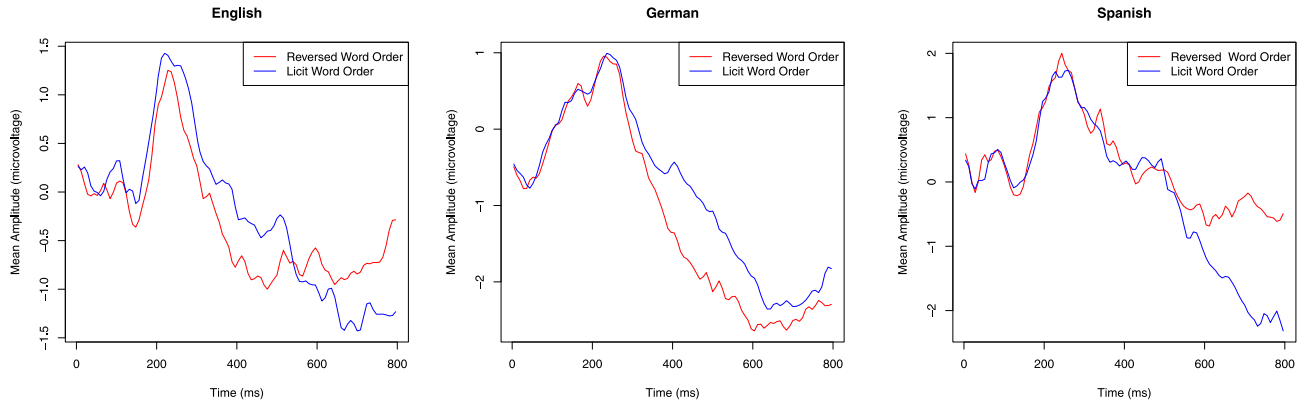


FIGURE A1 | Grand average ERP for the raw data at electrode C3, by participant group and Word Order condition.

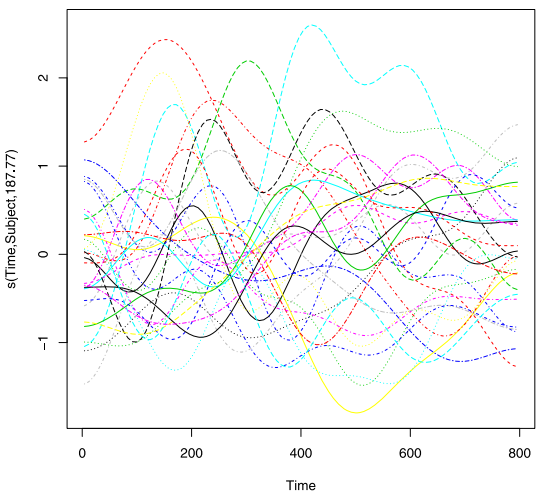


FIGURE A2 | Average amplitude (μV) over time (ms) by participant at electrode C3. Each line corresponds to a participant.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 13 January 2015
doi: 10.3389/fpsyg.2014.01525

Interface strategies in monolingual and end-state L2 Spanish grammars are not that different

María C. Parafita Couto[1]*, Virginia C. Mueller Gathercole[2] and Hans Stadthagen-González[3]

[1] *Leiden University Center for Linguistics and Leiden Institute for Brain and Cognition, Leiden University, Leiden, Netherlands*
[2] *Linguistics Program, English Department, Florida International University, Miami, FL, USA*
[3] *Department of Psychology, University of Southern Mississippi, Hattiesburg, MS, USA*

Edited by:
Vicky Chondrogianni, University of Edinburgh, UK

Reviewed by:
Marcel Den Dikken, Graduate Center of the City University of New York, USA
Aritz Irurtzun, Centre National de la Recherche Scientifique, France

****Correspondence:***
María C. Parafita Couto, Leiden University Center for Linguistics, Leiden University, van Wijkplaats 3 Room 005B, 2311 BX, Leiden, Netherlands
e-mail: m.parafita.couto@hum.leidenuniv.nl

This study explores syntactic, pragmatic, and lexical influences on adherence to SV and VS orders in native and fluent L2 speakers of Spanish. A judgment task examined 20 native monolingual and 20 longstanding L2 bilingual Spanish speakers' acceptance of SV and VS structures. Seventy-six distinct verbs were tested under a combination of syntactic and pragmatic constraints. Our findings challenge the hypothesis that internal interfaces are acquired more easily than external interfaces (Sorace, 2005, 2011; Sorace and Filiaci, 2006; White, 2006). Additional findings are that (a) bilinguals' judgments are less firm overall than monolinguals' (i.e., monolinguals are more likely to give extreme "yes" or "no" judgments) and (b) individual verbs do not necessarily behave as predicted under standard definitions of unaccusatives and unergatives. Correlations of the patterns found in the data with verb frequencies suggest that usage-based accounts of grammatical knowledge could help provide insight into speakers' knowledge of these constructs.

Keywords: interfaces, focus, unaccusative, unergative, L2 acquisition, Spanish

INTRODUCTION

This article concerns the extent to which high-functioning L2 Spanish speakers have acquired the full grammar for the expression of focus. In particular, we look at the language-specific means of expressing linguistic focus within the bilingual's two grammars, Spanish and English. Linguistic focus concerns that portion of a sentence that contributes the most relevant new information, the non-presupposed information, of the utterance. As such, focus stands at the interface between syntax, phonology, and pragmatics, as what is focused in a sentence, expressed syntactically and/or phonologically, depends directly on the discourse and pragmatic intent in which the sentence is embedded. The main questions explored in this article are (i) whether bilinguals' grammars converge with (or diverge from) those of monolinguals and (ii) whether certain linguistic areas are more vulnerable to influence than others. We will particularly examine the expression of focus in Spanish through the syntactic operations of word order. We test both Spanish monolinguals and Spanish-English functional bilinguals.

Recent research in linguistic theory has focused on the properties that (external) interface conditions impose on the design of the language faculty (Chomsky, 2005), since the output of the computational system has to be interpreted by other cognitive systems (sensory-motor systems and conceptual-intentional systems). L2 research has recently posed the question of how well L2 learners are able to integrate linguistic phenomena pertaining to interfaces (White, 2009). Sorace (2005), Sorace and Filiaci (2006) and Tsimpli and Sorace (2006) have formulated the interface hypothesis, which argues that phenomena contained within narrow syntax or lying at internal interfaces can be completely acquired in the L2, whereas full acquisition may not be possible for phenomena placed at external interfaces. These authors claim that narrow syntax is not a problem for acquisition while internal interfaces (at least syntax/semantics) are argued to be relatively unproblematic. However, external interfaces (e.g. syntax/discourse) are claimed to be a locus of instability in bilingual speakers. More recently, Sorace (2011) emphasized that the interface hypothesis predicts that both syntactic and pragmatic conditions are acquirable but the integration of both conditions remains less than optimally efficient, giving rise to optionality.

In this paper we explore a phenomenon that lies at both the external and internal interfaces: the expression of focus (syntax-discourse interface) in sentences with intransitive (unaccusative and unergative) verbs (syntax-semantics interface). We do this by exploring whether Spanish-English functional bilinguals have problems in coordinating the syntax and the pragmatics in focus contexts through the distribution of subject-verb (SV) and verb-subject (VS) word order. Spanish has flexible word order, while English is more rigid. While in neutral focus contexts SV is the canonical word order in Spanish, VS order can result from different kinds of syntactic operations (Lozano, 2003). According to Contreras (1978), Suñer (1982) and Zubizarreta (1998), there is a clear tendency for speakers to produce VS order for unaccusative verbs, and the most common discourse-neutral order for unergatives is SV. The intransitive verb class is of interest because of the contrast between Spanish and English. English allows stress on the preverbal subject for both unaccusative and unergative verbs in "out of the blue" contexts, in which the subject is the focus (e.g., "A book fell," cf. Schmerling, 1976; Selkirk, 1984; Nava, 2007). In the same context in Spanish, the subject would occur

in post-verbal position and stress would fall on the rightmost constituent ("Se cayó un libro").

This paper examines Spanish-English bilinguals' knowledge of the Spanish forms. The paper is structured as follows: First, we summarize the effects of unaccusativity and focus on word order in Spanish, report on previous research on the acquisition of Spanish word order patterns, and present the research questions and hypotheses. We then report on the experimental evidence bearing on these questions, followed by a discussion of our findings in relation to the perspective of previous research.

THEORETICAL BACKGROUND: IS WORD ORDER AT THE INTERFACES ACQUIRABLE?

RESEARCH QUESTION

Spanish and English are both SVO languages, but VS is also possible in both languages:

(1) Llegaron los niños.
Arrived the kids
"The kids arrived"
(2) Here comes the sun.

However, in Spanish post-verbal subjects seem to be produced freely with all verb classes:

(3) Ha telefoneado María al presidente. (transitive)
has phoned Mary the president
"Mary has phoned the president."
(4) Ha hablado Juan. (unergative)
has spoken Juan
"Juan has spoken."
(5) Ha llegado Juan (unaccusative)
has arrived Juan
"Juan has arrived."

In Spanish, inversion is usually a means of "focalization": preverbal subjects are topics (given information) and post-verbal subjects are focus (new information) (Zubizarreta, 1998; Belletti, 2001, 2004). e.g.,

(6) ¿'Quién ha llegado/hablado?
Who has arrived/spoken?

i. Ha llegado/hablado Juan
ii. #Juan ha llegado/hablado

In neutral (non-focus) contexts, subjects tend to be discourse-initial, except in the case of unaccusative verbs:

(7) a. Una mujer gritó (unerg)
b. # Gritó una mujer.
'A woman shouted.'
(8) a. # Una mujer llegó. (unacc)
b. Llegó una mujer.
'A woman arrived.'

Previous studies on Spanish native speakers show that verb choice may determine word order (Pinto, 1999; Hertel, 2003; Lozano, 2003, 2006a,b). Default word order is reported to be SV for unergatives and VS for unaccusatives (i.e., determined by the lexicon-syntax interface). Word order in focused contexts is VS for both verb types (i.e., determined by the syntax-discourse interface).

Unaccusativity: syntax or semantics?

Baker (1983) remarked that "all seemingly intransitive verbs are not created equal" (p. 1). According to the Unaccusative Hypothesis, there are two classes of intransitive verbs: unaccusatives and unergatives (Perlmutter, 1978; Perlmutter and Postal, 1984). For some researchers, the difference between the two types is semantic; for others, it is syntactic. Semantically, the two types of verb differ in that, whereas the subject of an unergative verb actively initiates or is actively responsible for the action expressed in the verb, the subject of an unaccusative verb does not. Subjects of unaccusatives bear the semantic role of theme or patient, usually associated with the objects of verbs. In Dowty's (1991) and van Valin's (1999) terms, the difference between the two classes of verbs reduces to differences in agentivity and telicity. Unergative verbs are typically agentive and denote an atelic process (run, walk, work), while unaccusative verbs (die, disappear, exist) are non-agentive and telic, usually denoting a change of some sort.

In contrast, for generative linguists such as Burzio (1986) and Rosen (1984) the distinction between the two classes of verbs is mainly syntactic. According to the Unaccusative Hypothesis, the single argument of unaccusatives is syntactically a direct object, while the single argument of unergatives is the subject. Thus, although superficially the sentences "The leaf fell" and "The bird chirped" both show NP-V word order, the former involves NP-movement from object to subject position (9), while in the latter the NP is base-generated in subject position (10).

(9) The leaf$_i$ fell t$_i$.
(10) The bird chirped.
(Friedmann et al., 2008)

Even though Burzio's (1986) formulation of the Unaccusative Hypothesis has been widely accepted, it is not uncontroversial. For example, Rappaport Hovav and Levin (2001:792) present counter-examples to syntactic accounts of English resultatives which are based on the assumption that result XPs are predicated of underlying direct objects. They concluded that "Our work calls even more seriously into question the existence of any evidence for the syntactic encoding of unaccusativity in English."

These different approaches have led to the result that in the literature, unaccusatives are not consistently classified in semantic or syntactic terms. Hatcher (1956) offered the first semantic classification. De Miguel (1993) took into account both theta-role structure and the semantics of the verb. Building on Burzio, Sorace (1995, 2000) considered both syntactic and semantic aspects in her classification. She proposed that there is a universal continuum (a "hierarchy") of gradients of

unaccusative/unergative verbs, a continuum of potentially universal significance. This hierarchy is based on the semantic concepts of telicity and agentivity. The extremes of the continuum ("core"), with non-agentive, telic meanings on one end and agentive, non-telic on the other, correspond to the prototypical unaccusative and unergative verbs. The verbs in the middle are more or less unaccusative or unergative, depending on where they lie on the continuum. The types of semantic meanings that fall between the two extremes are as shown in **Figure 1**.

Crosslinguistically, unaccusative verbs fall on one end and unergative verbs on the other, and the two categories of verbs are distinguished by differences in syntactic behavior. Languages differ, however, in terms of the point at which unaccusatives are separated from unergatives along the hierarchy. But Sorace and Shomura (2001) raise the issue of the learnability of the unaccusative/unergative dichotomy and posit that the difficulty in acquiring this split intransitivity (unaccusatives vs. unergatives) stems from the problem of systematically linking "a multicategorial lexical-semantic level to a necessarily binary syntactic level…" (p. 249).

Unaccusativity and learnability

Montrul (2001) points out that for linguists and psycholinguists working within the generative framework (Chomsky, 1981, 1995), who assume that there is a syntactic difference between the two classes of intransitive verbs, the acquisition of unaccusativity represents a classic "poverty of the stimulus" problem. On the surface, all intransitive verbs look alike: they have one argument. How does the learner find out, solely from positive evidence, that these two verb classes have different underlying representations? Furthermore, when the learner finally finds out that there is a distinction, how does he/she classify newly acquired intransitive verbs? van Hout (1996) argues that the L1 learner already comes equipped with knowledge of the syntactic distinction (i.e., it is innate) but needs to find out which specific semantic notion is grammatically relevant for the unaccusative/unergative classification (telicity, change of state, transition, etc.).

Available studies report very few problems with the acquisition of intransitive verbs in L1 acquisition (e.g., van Hout et al., 1992 for Dutch). As argued by Montrul (2001), one of the main differences between child language acquisition and adults acquiring a second language is that second language (L2) learners already have a mature linguistic system in place. Therefore, if unaccusativity is universal, L2 learners presumably know about the unaccusative/unergative distinction (and know how it is expressed in their native language), although the semantic basis for the distinction might be different in the L1 and the L2. In English, certain unaccusative verbs can appear with existential subjects ("There appeared three men.") and in the resultative construction ("The bag fell open.") whereas unergative verbs cannot ("∗There worked three men." "∗Mary laughed hoarse.") (Perlmutter, 1978; Levin and Rappaport Hovav, 1995; Montrul, 2004). However, quite differently from what has been reported for L1 acquisition, a number of L2 acquisition studies have reported that unaccusative verbs, but not unergatives, cause problems for L2 learners of English and other languages and of various L1 backgrounds, especially at high intermediate and quite advanced levels (e.g., Yip, 1995; Oshita, 2001; Sorace and Shomura, 2001, inter multa al.). It has been reported that L2 learners have difficulty in determining the range of appropriate syntactic realizations of the distinction and that this difficulty can persist into near-native levels of proficiency (Hawkins, 2000).

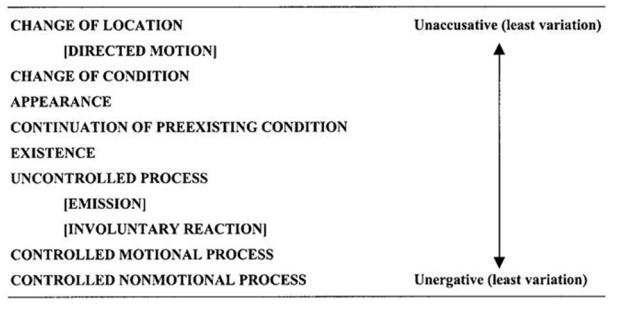


FIGURE 1 | Sorace's unaccusativity hierarchy (Sorace and Shomura, 2001).

Montrul (2005) rightly argued that Spanish is an interesting testing ground because, unlike Italian, which has auxiliary selection and *ne*-cliticization, Spanish does not provide such robust and clear syntactic and morphological evidence for unaccusativity. Furthermore, the topic has remained largely understudied in Spanish L2 acquisition. For example, Montrul points out that it is not known what role, if any, semantic subclass plays in the acquisition of these verbs. One very recent study (de Prada Pérez and Pascual y Cabo, 2012), however, suggests that even though Spanish heritage speakers use subject position differently in broad and narrow focus, they make no distinction between unergative and unaccusative predicates (contra the predictions of the Interface Hypothesis).

We will test below the assumption that unaccusativity corresponds to a syntactic phenomenon related to word order in Spanish and explore the possibility that semantics also plays a role in the classification of verbs, as has been indicated for other Romance languages (Sorace, 1993a,b, 1995). Before addressing these issues, however, it is essential to examine another factor that also constrains the distribution of SV/VS order in Spanish: Focus.

SV/VS: unaccusativity and/or focus

Lozano's (2003) dissertation was perhaps the first attempt to argue that the distribution of SV and VS in L2 Spanish is constrained both by universal principles like the Unaccusative Hypothesis and, at the same time, by discourse parameterizable features like presentational focus. He argues that learners' knowledge is convergent in unaccusative/unergative contexts (internal interface) yet divergent in presentational focus contexts (external interface). A few studies on the acquisition of the syntax-discourse interface previous to Lozano's dissertation had reported that presentationally focused subjects in final position are acquired late in L2 Spanish—e.g., Hertel (2003), as was also reported for L2 Italian (Belletti and Leonini, 2004). Ocampo (1990) and Camacho (1999) similarly reported that the acquisition of distinct word orders to mark focus in Spanish is acquired late or perhaps never in native-like fashion. More recently, Domínguez and

Arche (2014) argued after looking at their native data that the linguistic evidence available for acquiring the syntactic properties of unergative and unaccusative verbs in Spanish is not completely transparent and that L2 speakers may not get clear evidence, which can explain why learners find the acquisition of SV–VS contrasts persistently difficult. However, they claim that their analysis is compatible with the view that L2 speakers eventually converge on the grammar of native speakers, and that this may well be the case with their advanced speakers as their experience in the L2 increases.

Research questions and hypotheses

In light of the contributions of syntax, semantics and information structure (i.e., Focus) to the acceptability and production of SV/VS word order in Spanish, the following research questions can be formulated with respect to L2 Spanish speakers who are long-standing functional[1] bilinguals:

(i) Do long-standing functional Spanish-English bilinguals (L1 English, L2 Spanish) respect syntactic differences between unaccusative and unergative verbs?
(ii) Does the hierarchy proposed by Sorace (2000) play a role in the acquisition/processing of these verbs in L2 Spanish?

In this study we set out to compare syntactic, pragmatic, and lexical influences on adherence to SV and VS orders in Spanish monolinguals and in fluent L2 speakers of Spanish. This study also looked for empirical evidence to test Beck's (1998) claim that optionality results in a permanent state even after long immersion in the language. We look at the distribution of SV/VS order in long-term Spanish L2 speakers and whether long experience with the L2 leads to convergent native-like behavior, as argued by Domínguez and Arche (2014). Currently, there is no consensus on the status of optionality in end-state grammars (Lozano, 2003, 2009). In addition, this study provides an empirical test for the interface hypothesis, which posits that internal interfaces (e.g., unaccusativity as a syntax-semantics interface) are not problematic for L2 acquisition while external interfaces (e.g., focus as a syntax-discourse interface) are problematic, acquired later, or never acquired. The following predictions were made:

– If syntactic knowledge of unaccusativity develops early, native speakers should have robust syntactic knowledge of the distinction between unaccusative and unergative verbs.
– If Sorace's Hierarchy is valid, we should observe differences in how monolinguals rate different semantic subclasses of unaccusative and unergative verbs.
– If there are differences between monolinguals and functional bilinguals, these should be observed in their ratings of the acceptability of unaccusative and unergative constructions.
– L2 speakers may have a less robust knowledge of the unaccusative/unergative distinction with non-prototypical verbs—i.e., verbs in the center of the continuum.
– If Lozano's claim regarding the ease with which L2 learners learn internal interfaces relative to learning external interfaces is correct, functional bilinguals should differ in their ratings from monolinguals more with regard to focus environments than with regard to verb class, given that focus lies at an external interface.

[1]Functional bilingualism is one's ability to use and produce both languages across "an encyclopedia of everyday events" (Baker, 1993:13).

THE EXPERIMENT

METHODS

We conducted an experimental study to test the knowledge of Spanish-English functional bilinguals in comparison with Spanish monolinguals with respect to unergative and unaccusative SV/VS alternations both in neutral and presentational focus contexts. For our study, we followed the methodology developed by Lozano (2003) but modified it to test a higher number [$n = 76$] of Spanish verbs classified according to Sorace's Unaccusative hierarchy.

Participants

A total of 40 subjects participated in the study. All but two of the participants lived in Madrid, Spain. A group of 20 Spanish monolingual native speakers (mean age 23, range 19–31) served as a baseline to compare with the L2 Spanish speakers' results. The experimental group consisted of 20 English native speakers aged 29–72 (mean age 46.6), eighteen of whom lived in Madrid, Spain, and had lived there for an average period of 20.7 years (range 3–47 years). The remaining two of the L2 Spanish speakers had lived for periods in Spain; one of them was a university psychology professor who had been married to a Spaniard for over 20 years, and the language of conversation in their home between themselves and with their children was Spanish, and the other was a college professor of Spanish literature who had been traveling to Spain on a regular basis since 1968 and used Spanish in her work environment on a daily basis. The onset of Spanish for the L2 bilinguals occurred between the ages of 16 and 38 (average age of onset: 21.4 years), and all speak Spanish on a daily basis at work and with their families. Three of the bilinguals were male, 16 female; 7 of the monolinguals were male, 13 female. The educational backgrounds of the bilingual participants was high school level or higher (4 high school, 12 BA/BSc or equivalent, 3 MA, PhD or equivalent). On a scale of 1 to 4 (1 = only some words, 2 = confident in basic conversations, 3 = fairly confident in extended conversations, 4 = confident in extended conversations), the bilinguals rated their English abilities at 3.95 (19 ratings of "4," 1 of "3") and their Spanish abilities at 3.74 (5 ratings of "3," 15 of "4"). All participants signed consent forms and were paid for their participation.

Stimuli

The instrument employed was an acceptability judgment test. Sentences were constructed with 76 distinct verbs embedded within short scenarios depicting a context of utterance. The target stimuli involved 19 unergative verbs in neutral contexts, 19 unaccusatives in neutral contexts, 19 unergatives in focused contexts, and 19 unaccusatives in focused contexts. The verbs tested included semantically prototypical, semantically non-prototypical, and semantically intermediate/less prototypical

verbs, according to Sorace's Hierarchy. The verbs also were grouped according to whether they typically occurred with or without multifunctional se. Sample verbs are shown in **Figure 2**.

Half of the conditions presented contexts appropriate for focused subjects and half for non-focused subjects. Each contextual setting ended with a question, followed by two possible replies (see **Figure 3**). The two possible replies represented different word orders (SV vs. VS).

As exemplified in **Figure 3**, each target sentence was accompanied by a 5-point Likert rating scale (see **Figure 4**). (Participants were given only the Spanish; the translation is provided here for the convenience of the reader.) Value 1 corresponded to *No se puede decir así* "you cannot say it like this," value 5 corresponded to *Está perfecto decirlo así* "it's perfect to say it like that," with values 2–4 several levels between these two extremes. (Value 0 corresponded to *No sé si se puede decir así* "I don't know if you can say it like this").

Twenty-four control sentences were also included. These sentences involved pro-drop, with the two choice answers differing in the presence/absence of overt subjects. Two training stimuli were placed at the beginning of the test. These consisted of structures not related to those of interest here, one trial involving the position of a clitic pre- or post-verbally and the other the order of a noun-adjective sequence.

Four randomized versions of the test were created with the same sentences but with different sequential order. The sequential order was randomized following Cowart (1997) "blocking" procedure.

Syntactically	Unergative (SV)		Unaccusative (VS)	
Semantically prototypical	**Controlled process (non-motional)**		**Change of location**	
	hablar	"speak"	llegar	"arrive"
	cantar	"sing"	salir	"leave"
	gritar	"shout"	venir	"come"
	protestar	"protest"	entrar	"enter"
Semantically less-prototypical	**Controlled process (motional)**		**Change of state**	
	correr	"run"	morir	"die"
	caminar	"walk"	caducar	"expire"
	bajar	"lower"	surgir	"arise"
	bailar	"dance"	desaparecer	"disappear"
Semantically non-prototypical	**Uncontrolled process (non-volitional)**		**Existence or state**	
	temblar	"tremble"	existir	"exist"
	toser	"cough"	quedar	"stay"
	bostezar	"yawn"	faltar	"be missing"
	roncar	"snore"	sobrar	"be left over"

FIGURE 2 | Sample verbs.

Trabajas en una guardería y Pablito empieza a llorar mucho porque otro niño, Diego, llegó a la guardería. Tu compañera de trabajo, María, sabe que Pablito siempre llora cuando llega alguien y te pregunta: ¿Quién llegó? Tú respondes:

(a) *Diego llegó.* 1 2 3 4 5 | 0

(b) *Llegó Diego.* 1 2 3 4 5 | 0

You work at a nursery and Pablito starts crying because another child, Diego, just arrived. Your workmate, Maria, knows that Pablito always cries when someone arrives and asks you: Who arrived? You answer:

(c) *Diego arrived.* 1 2 3 4 5 | 0

(d) *Arrived Diego.* 1 2 3 4 5 | 0

FIGURE 3 | Sample stimulus.

Procedure

Participants were asked to judge the acceptability of both sentences given. Following Lozano's model (Lozano, 2003, 2006a[2]), participants were given written instructions at the beginning of the test. The instructions highlighted that the researchers were interested in the participant's opinion of a set of sentences, as follows:

> "El objetivo de este test es averiguar cómo te suenan ciertas oraciones en español. Es importante resaltar que sólo nos interesa TU opinión sobre ellas, es decir, si te parecen más o menos aceptables. El test no será corregido, sino que su finalidad es averiguar si ciertas oraciones suenan mejor o peor a los hablantes nativos de español."

English translation (provided here for the convenience of the reader; the English version was not given to the participants):

> "The objective of this test is to see how certain sentences sound to you in Spanish. It is important to stress that we are only

[2]Lozano's instrument can be downloaded from the IRIS database (http://www.iris-database.org)

No se puede decir así.	Me parece mal pero no estoy seguro-a.	No me parece muy mal, pero no me parece la mejor manera de decirlo.	Está más o menos bien.	Está perfecto decirlo así.	No sé si se puede decir así.
1	2	3	4	5	0
You cannot say it like this.	I think it's bad, but I'm not sure.	I don't think it is very bad, but I don't think it's the best way to say it.	It's more or less OK.	It's perfect to say it like that.	I don't know if you can say it like this.
1	2	3	4	5	0

FIGURE 4 | Rating scale.

interested in YOUR opinion about them—that is, if they seem more or less acceptable to you. This test will not be graded, but the goal is to see if certain sentences sound better or worse to those who speak Spanish natively."

It also contained explicit instructions on how to complete the test and it detailed what the value scale meant, providing some examples, as follows:

"En cada número que sigue, verás una lectura corta. Léela primero. Luego le siguen dos oraciones muy parecidas. Oración (a) y oración (b). Queremos que juzgues, dada la lectura que acabas de leer, cómo suena cada oración. Cada una de las oraciones está seguida de la siguiente escala para puntuar cada oración:

No se Puede decir así	Me parece mal pero no estoy seguro-a	No me parece muy mal, pero no me parece la mejor manera de decirlo	Está más menos bien	Est'a perfecto decirlo as'i	No sé si se puede decir así
1	2	3	4	5	6

Aquí te ponemos un ejemplo:
A ti siempre te gustaron mucho los churros con chocolate. Cuando eras pequeño-a, siempre que veías churros, le decías a tu madre:

(a) Quiero comerlos. 1 2 3 (5) 0
(b) Los quiero comer. 1 2 3 (5) 0

English translation (given here for the reader):

For every item below, you will see a short reading. Read it first. After each item, there are two very similar sentences, sentence (a) and sentence (b). Please judge, based on the reading you have just read, how each sentence sounds. Each of the sentences is followed by the following scale for you to rate each sentence.

You cannot say it like this	I think it's bad, but I'm not sure	I don't think it is very bad, but I don't think it's the best way to say it	It's more or less OK	It's perfect to say it like that	I don't know if you can say it like this
1	2	3	4	5	0

Here is an example:

You always really liked churros with chocolate. When you were young, whenever you saw churros, you would tell your mother:

(a) I want to eat them. 1 2 3 4 (5) 0
(b) Them I want to eat. 1 2 3 4 (5) 0

The test also emphasized that any combination of numbers was possible [i.e., sentence (a) could be 5 and sentence (b) could be 1, or both of them could be 5, etc.]. Subjects were asked to do the test as quickly as possible, as we were only interested in their first intuitions.

RESULTS

General results

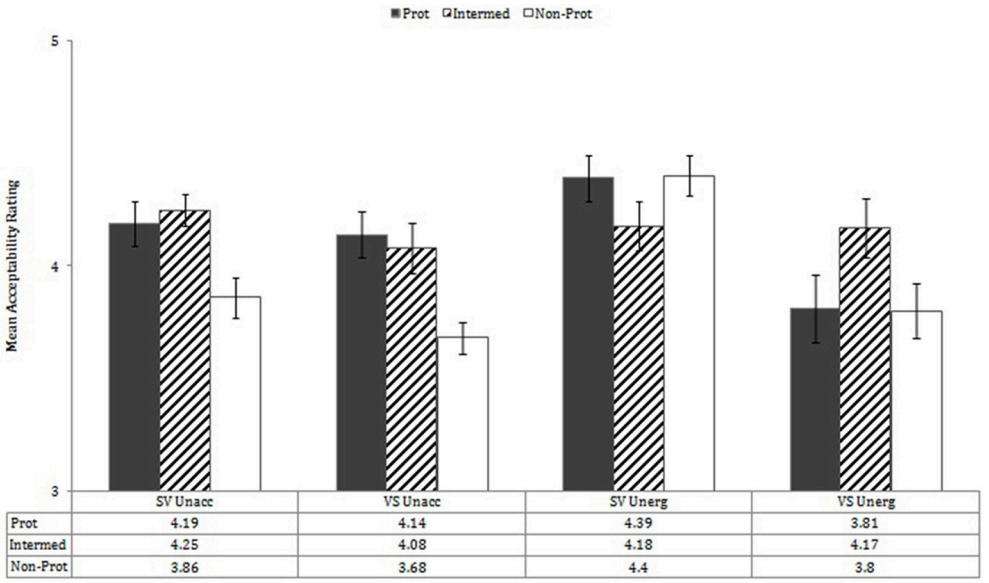
An Five-Way mixed repeated measures ANOVA was conducted in which verb type (unaccusative, unergative), word order (SV, VS), prototypicality (3 levels), information structure (focus, non-focus), and participant group (monolingual, bilingual) were entered as variables. Results revealed, first, main effects of verb type, $F_{(1,\ 38)} = 5.29$, $p = 0.027$, $\eta^2 = 0.122$, word order, $F_{(1,\ 38)} = 7.71$, $p = 0.008$, $\eta^2 = 0.169$ and prototypicality, $F_{(2,\ 76)} = 21.25$, $p < 0.001$, $\eta^2 = 0.359$. Unergative sentences tended to receive higher scores (4.12, $SEM = 0.076$) than unaccusative sentences (4.03, $SEM = 0.064$); sentences with SV order received higher acceptability scores overall (4.20, $SEM = 0.074$) than those with VS order (3.95, $SEM = 0.090$); and sentences with prototypical and intermediate verbs received higher acceptability scores overall (prototypical 4.13, $SEM = 0.075$, intermediate 4.17, $SEM = 0.073$) than those with non-prototypical verbs (3.93, $SEM = 0.066$), pairwise comparisons $p < 0.001$.

These main effects were modified, however, by interaction effects. First, there were interactions of Verb Type X Prototypicality, $F_{(2,\ 76)} = 12.07$ $p < 0.001$, $\eta^2 = 0.241$; Verb Type X Word Order, $F_{(1,\ 38)} = 8.06$, $p = 0.007$, $\eta^2 = 0.175$; Prototypicality X Word Order, $F_{(2,\ 76)} = 4.98$, $p = 0.009$, $\eta^2 = 0.116$; and Verb Type X Prototypicality X Word Order, $F_{(2,\ 76)} = 3.98$, $p = 0.023$, $\eta^2 = 0.095$. Performance by verb type, prototypicality, and word order is shown in **Figure 5**.

Follow-up analyses in which each verb type was analyzed separately via mixed-effect repeated measures (with word order (SV, VS), prototypicality (3 levels), information structure (focus, non-focus), and participant group (monolingual, bilingual) as variables) revealed that for unaccusatives, performance by word order was not significant, nor was performance by Word Order X Prototypicality; there was, however, a significant overall effect of prototypicality, $F_{(2,\ 76)} = 40.29$, $p < 0.001$, with sentences involving non-prototypical verbs judged less acceptable than those with prototypical or intermediate verbs, ps < 0.001.

For unergatives, in contrast, judgments varied by word order, $F_{(1,\ 38)} = 10.12$, $p = 0.003$, and by Word Order X Prototypicality, $F_{(2,\ 76)} = 6.94$, $p = 0.002$. For SV unergative sentences, prototypicality was significant, $F_{(2,\ 76)} = 3.28$, $p = 0.043$, but only in relation to significantly higher acceptability ratings of SV with non-prototypical verbs than with intermediate verbs ($p = 0.029$) (and near-significantly higher with prototypical than intermediate, $p = 0.083$). For VS unergative sentences, judgments showed a reverse pattern: Prototypicality was significant, $F_{(2,\ 76)} = 5.90$, $p = 0.004$, but VS sentences built on intermediate verbs received higher scores than those built on either prototypical or non-prototypical verbs, $p = 0.008$, $p = 0.011$, respectively.

Thus, for unaccusatives, word order in general did not affect performance (but see results concerning information structure, below), and sentences built on non-prototypical verbs were judged less acceptable than those in which prototypical and intermediate verbs were used. For unergatives, in contrast, SV sentences were in general judged acceptable (but less so for the intermediate verbs), and VS sentences were judged less acceptable, especially in cases in which prototypical and non-prototypical verbs occurred.



	SV Unacc	VS Unacc	SV Unerg	VS Unerg
Prot	4.19	4.14	4.39	3.81
Intermed	4.25	4.08	4.18	4.17
Non-Prot	3.86	3.68	4.4	3.8

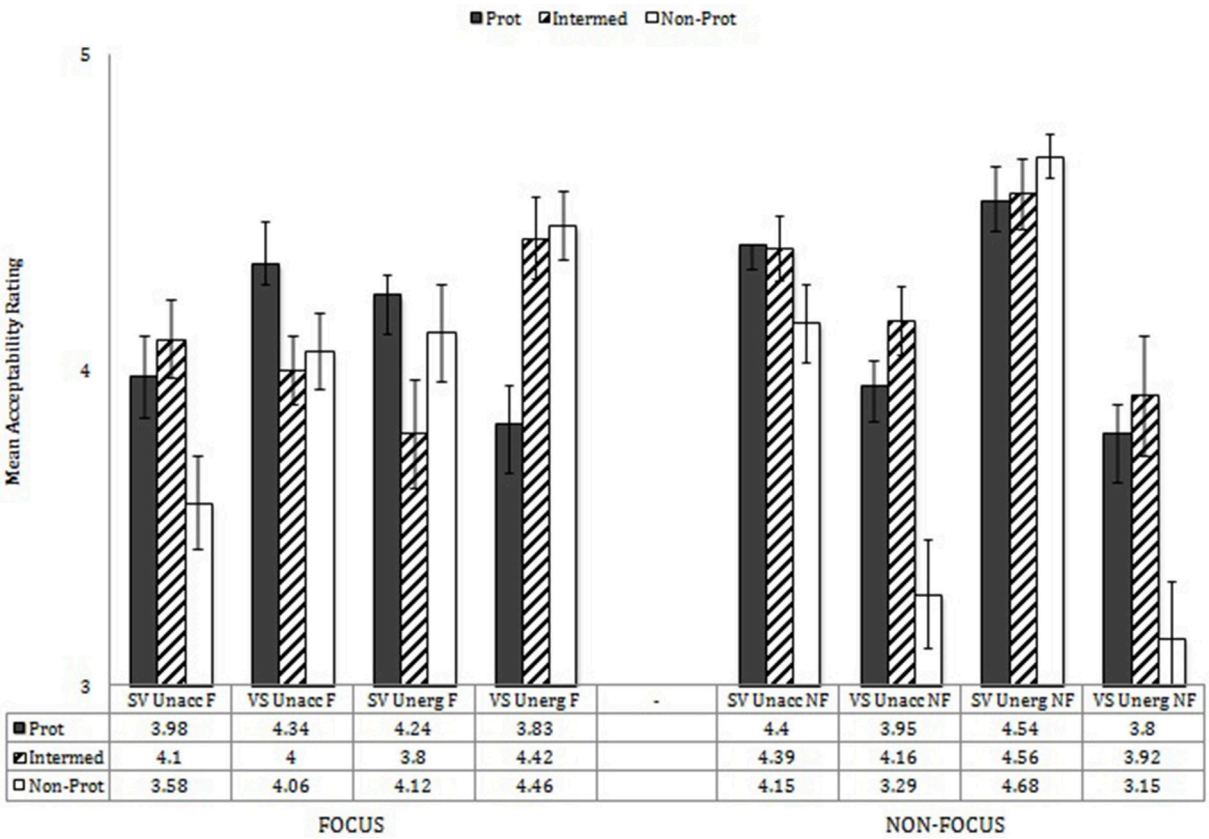
FIGURE 5 | Mean acceptability ratings by word order X verb type X prototypicality. Error bars in all figures show SEM.

The main analyses also showed significant interactions of Word Order X Information Structure, $F_{(1, 38)} = 47.39$, $p < 0.001$, $\eta^2 = 0.555$; Prototypicality X Information Structure, $F_{(2, 76)} = 13.72$, $p < 0.001$, $\eta^2 = 0.265$; Verb Type X Prototypicality X Information Structure, $F_{(2, 76)} = 4.03$, $p = 0.022$, $\eta^2 = 0.092$; Prototypicality X Word Order X Information Structure, $F_{(2, 76)} = 8.59$, $p < 0.001$, $\eta^2 = 0.184$; and Verb Type X Prototypicality X Word Order X Information Structure, $F_{(2, 76)} = 6.53$, $p = 0.002$, $\eta^2 = 0.147$. Performance by verb type, prototypicality, word order, and information structure is shown in **Figure 6**. These effects were explored, first, by analysing each verb type separately.

For unaccusative verbs, there were interactions of Prototypicality X Information Structure, $F_{(2, 76)} = 8.26$, $p = 0.001$; Word Order X Information Structure, $F_{(1, 38)} = 24.05$, $p < 0.001$; and Prototypicality X Word Order X Information Structure, $F_{(2, 76)} = 4.61$, $p = 0.013$. Unaccusatives with SV word order showed significant differences in acceptability by prototypicality, $F_{(2, 76)} = 12.35$, $p < 0.001$: SV was more accepted with prototypical and intermediate verbs than with non-prototypical verbs, $p = 0.001$, $p < 0.001$, respectively. Unaccusatives with VS word order showed significant effects of prototypicality, $F_{(2, 76)} = 14.54$, $p < 0.001$, information structure, $F_{(1, 38)} = 15.68$, $p < 0.001$, and of Prototypicality X Information Structure, $F_{(2, 76)} = 10.67$, $p < 0.001$. In focus contexts, acceptability of unaccusatives with VS order showed significant effects by prototypicality, $F_{(2, 76)} = 4.83$, $p = 0.011$, with higher acceptability ratings with prototypical verbs than with either intermediate or non-prototypical verbs, $p = 0.002$, $p = 0.035$, respectively. In non-focus contexts, acceptability of unaccusatives with VS order differed across the three prototype levels, $F_{(2, 76)} = 16.38$, $p < 0.001$: prototypical less than intermediate $p = 0.018$; prototypical greater than non-prototypical $p < 0.001$; intermediate greater than non-prototypical $p < 0.001$. Interestingly, these results show higher acceptability ratings of VS in non-focus contexts with intermediate unaccusative verbs than with prototypical unaccusatives.

For unergative verbs, there was a main effect of word order, $F_{(1, 38)} = 10.12$, $p = 0.003$, and there were significant interactions of Prototypicality X Word Order, $F_{(2, 76)} = 6.94$, $p = 0.002$; Prototypicality X Information Structure, $F_{(2, 76)} = 10.34$, $p < 0.001$; Word Order X Information Structure, $F_{(1, 38)} = 36.63$, $p < 0.001$, and Prototypicality X Word Order X Information Structure, $F_{(2, 76)} = 13.23$, $p < 0.001$. When unergatives occurred with SV word order, effects of prototypicality, $F_{(2, 76)} = 3.28$, $p = 0.043$, informational structure, $F_{(1, 38)} = 24.70$, $p < 0.001$, and of Prototypicality X Informational Structure, $F_{(2, 76)} = 3.28$, $p = 0.043$, reveal that whereas in non-focus contexts, there was no difference by prototypicality (note that SV structures with unergatives in non-focus contexts received the highest acceptability ratings out of all groups), in focus contexts, prototypicality effects were evident, $F_{(2, 76)} = 4.34$, $p = 0.016$, with higher acceptability ratings in relation to prototypical and non-prototypical verbs than with intermediate verbs, $p = 0.013$, $p = 0.048$, respectively. When unergatives were used with VS word order, in focus contexts, significant effects of prototypicality, $F_{(2, 76)} = 12.02$, $p < 0.001$, indicate lower ratings with prototypical verbs than with intermediate or non-prototypical verbs, $p = 0.001$, $p < 0.001$, respectively. When unergatives occurred with VS order in non-focus contexts, significant effects of prototypicality, $F_{(2, 76)} = 11.05$, $p < 0.001$, indicate significantly lower ratings with non-prototypical verbs than with either prototypical or intermediate verbs, $p < 0.001$, $p = 0.001$, respectively.

These results indicate that in non-focus contexts, SV order is accepted (in fact, preferred) with all verb types. Further, in non-focus contexts, VS order is more accepted with prototypical and



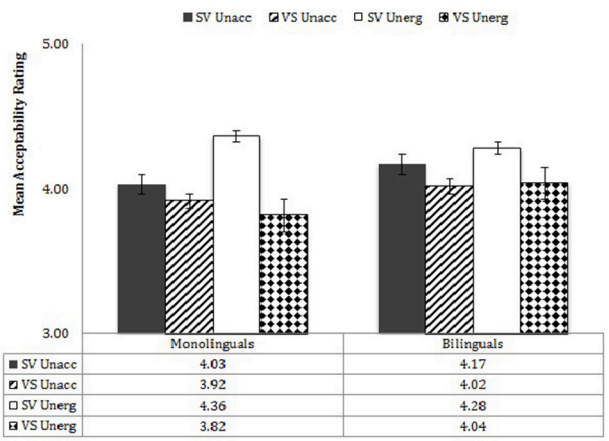
	SV Unacc F	VS Unacc F	SV Unerg F	VS Unerg F	-	SV Unacc NF	VS Unacc NF	SV Unerg NF	VS Unerg NF
Prot	3.98	4.34	4.24	3.83		4.4	3.95	4.54	3.8
Intermed	4.1	4	3.8	4.42		4.39	4.16	4.56	3.92
Non-Prot	3.58	4.06	4.12	4.46		4.15	3.29	4.68	3.15

FIGURE 6 | Performance by verb type, prototypicality, word order, and information structure.

intermediate verbs of both types (unergative and unaccusative) than with non-prototypical verbs.

In focus contexts, the results are more complex: With unaccusative verbs, VS order is preferred for prototypical and non-prototypical verbs, but with intermediate verbs, SV and VS are equally accepted. With unergative verbs, VS is preferred with intermediate and non-prototypical verbs, but SV is preferred with prototypical verbs.

These overall results are sometimes consistent with word order predictions regarding unaccusative and unergative verbs, sometimes inconsistent. Consistent with predictions, all unergatives in non-focus contexts are accepted with SV word order, and prototypical and intermediate unaccusative verbs are accepted with VS order. And in focus contexts, prototypical unaccusatives are accepted with VS order; however, prototypical unergatives are disfavored with VS order. Inconsistent with predictions, however, are the following: In non-focus contexts all types of unaccusative verbs are judged acceptable with SV order, and in focus contexts, prototypical unergative verbs are judged acceptable, in fact preferred, with SV word order.



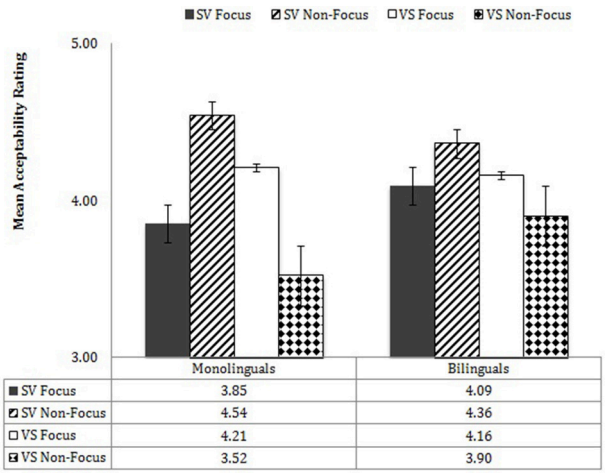
	Monolinguals	Bilinguals
SV Unacc	4.03	4.17
VS Unacc	3.92	4.02
SV Unerg	4.36	4.28
VS Unerg	3.82	4.04

FIGURE 7 | Verb type X WO X participant group.

Participant groups

Returning to the main analyses, let us now examine effects concerning participant groups. There was a near-significant effect of Verb Type X Word Order X Participant Group, $F_{(1,\ 38)} = 3.44$, $p = 0.071$, $\eta^2 = 0.083$, a significant interaction of Word Order X Information Structure X Participant Group, $F_{(1,\ 38)} = 9.47$, $p = 0.004$, $\eta^2 = 0.199$, and a near-significant interaction of Prototypicality X Word Order X Information Structure X Participant Group, $F_{(2,\ 76)} = 2.82$, $p = 0.066$, $\eta^2 = 0.069$.

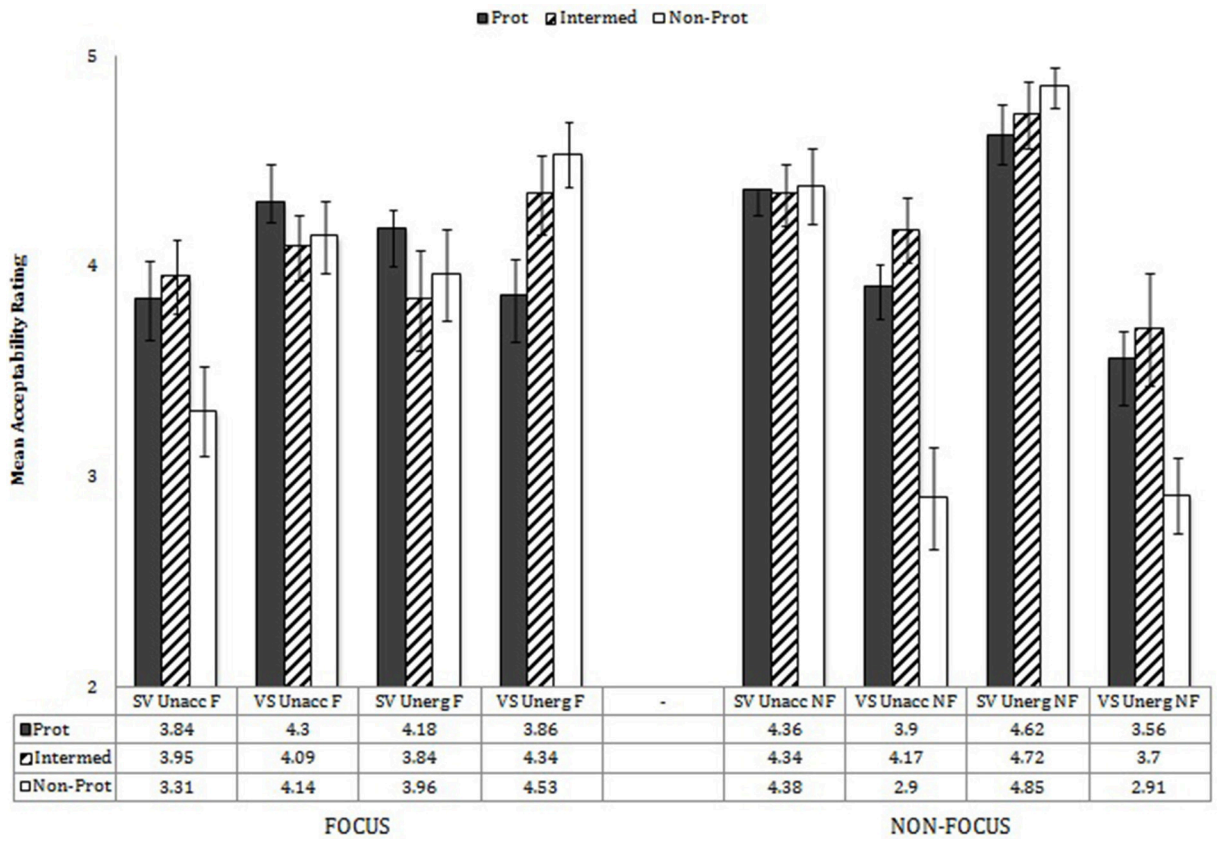
To explore these interactions, performance of the monolinguals and bilinguals was analyzed separately in mixed effects analyses with verb type (unaccusative, unergative), word order (SV, VS), prototypicality (3 levels), and information structure (focus, non-focus) as variables. Performance by each group by verb type and word order is shown in **Figure 7**. ANOVAs for the two separate participant groups showed that the bilinguals showed no main effect of either verb type or word order, nor an interaction of Verb Type X Word Order, whereas monolinguals showed significant effects for all three—verb type, $F_{(1,\ 19)} = 5.70$, $p = 0.03$, $\eta^2 = 0.231$, word order, $F_{(1,\ 19)} = 6.55$, $p = 0.019$, $\eta^2 = 0.256$, Verb Type X Word Order, $F_{(1,\ 19)} = 5.60$, $p = 0.007$, $\eta^2 = 0.329$. This means that, while the monolinguals distinguished the privileges of occurrence of the two types of verbs relative to word order, the bilinguals did not make any distinction between the two verb types and accepted both word orders about equally (but see below with regard to non-focus contexts).



	Monolinguals	Bilinguals
SV Focus	3.85	4.09
SV Non-Focus	4.54	4.36
VS Focus	4.21	4.16
VS Non-Focus	3.52	3.90

FIGURE 8 | Word order X information structure X participant group.

Monolinguals' and bilinguals' performance by Word Order X Information Structure is shown in **Figure 8**. Bilinguals showed a significant interaction of Word Order X Information Structure, $F_{(1,\ 19)} = 13.44$, $p = 0.002$, $\eta^2 = 0.419$, as did the monolinguals, $F_{(1,\ 19)} = 33.96$, $p < 0.001$, $\eta^2 = 0.641$. In the case of the bilinguals, in the Focus contexts, there was no significant difference in judgments for SV vs. VS word order [$F_{(1,\ 19)} = 0.22$, $p = 0.643$]; in the Non-Focus contexts, there was a significantly higher acceptance of SV order (4.36) than VS order (3.90). In the case of the monolinguals, in the Focus contexts, there was a near-significant preference for VS order (4.21) over SV order (3.85), $F_{(1,\ 19)} = 3.47$, $p = 0.078$; in the Non-Focus contexts, there was a dramatic preference for SV order (4.54) over VS order (3.52), $F_{(1,\ 19)} =$



	SV Unacc F	VS Unacc F	SV Unerg F	VS Unerg F	-	SV Unacc NF	VS Unacc NF	SV Unerg NF	VS Unerg NF
Prot	3.84	4.3	4.18	3.86		4.36	3.9	4.62	3.56
Intermed	3.95	4.09	3.84	4.34		4.34	4.17	4.72	3.7
Non-Prot	3.31	4.14	3.96	4.53		4.38	2.9	4.85	2.91

FOCUS NON-FOCUS

FIGURE 9 | Monolinguals' performance by verb type X word order X focus X prototypicality.

44.95, $p < 0.001$. Thus, the greatest differences between the two participant groups are that monolinguals showed a preference for VS order in Focus contexts, whereas bilinguals did not differentiate orders in those contexts, and the monolinguals showed a more dramatic categorical choice of SV over VS in non-focus contexts than the bilinguals, even though the latter also showed a significant difference in the acceptability of SV over VS in non-focus contexts.

To explore the near-significant interaction of Prototypicality X Word Order X Information Structure X Participant Group, the two participant groups' performance was compared for the three prototypical levels in a Two-Way mixed ANOVA (with prototype level and participant group as variables) for each verb type in each condition—focus/non-focus and SV/VS. Performance of the monolinguals by Prototypicality X Word Order X Information Structure is shown in **Figure 9** and of the bilinguals in **Figure 10**. Comparisons showed that the only significant differences in performance patterns for the two participant groups relative to prototypicality, word order, and information structure occurred with unaccusative verbs in the non-focus conditions, for both SV and VS order. With SV order for unaccusatives in the non-focus condition, there was an interaction between Prototypicality X Participant Group, $F_{(2,\ 76)} = 3.23$, $p = 0.045$: While the monolinguals treated all prototypical levels equivalently here (accepting SV for all prototypicality levels), the bilinguals showed lower acceptability ratings for the constructions with the non-prototypical verbs here than with prototypical or intermediate verbs, $p = 0.007$, $p = 0.006$, respectively. With VS order for unaccusatives in the non-focus condition, an interaction between Prototypicality X Participant Group, $F_{(2,\ 76)} = 3.68$, $p = 0.03$, revealed that while the bilinguals showed no significant difference in performance by prototypicality, the monolinguals found the constructions with non-prototypical verbs here to be less acceptable than those with the prototypical and intermediate verbs, p's < 0.001, and those with the prototypical verbs marginally less acceptable than those with intermediate verbs, $p = 0.07$. Thus, in general, the patterns of responses relative to the three prototypicality levels within each condition were similar for monolinguals and bilinguals. The exceptions were the following: (1) In the case of unaccusatives in non-focus contexts, monolinguals found SV order uniformly acceptable, while bilinguals tended to accept SV with non-prototypical verbs less than with prototypical and intermediate verbs. (2) In non-focus contexts, bilinguals treated VS orders as equally acceptable with unaccusatives at all prototypicality levels, but monolinguals tended to disallow VS with non-prototypical unaccusative verbs.

DISCUSSION

These results overall indicate the following:

With regard to the general findings:

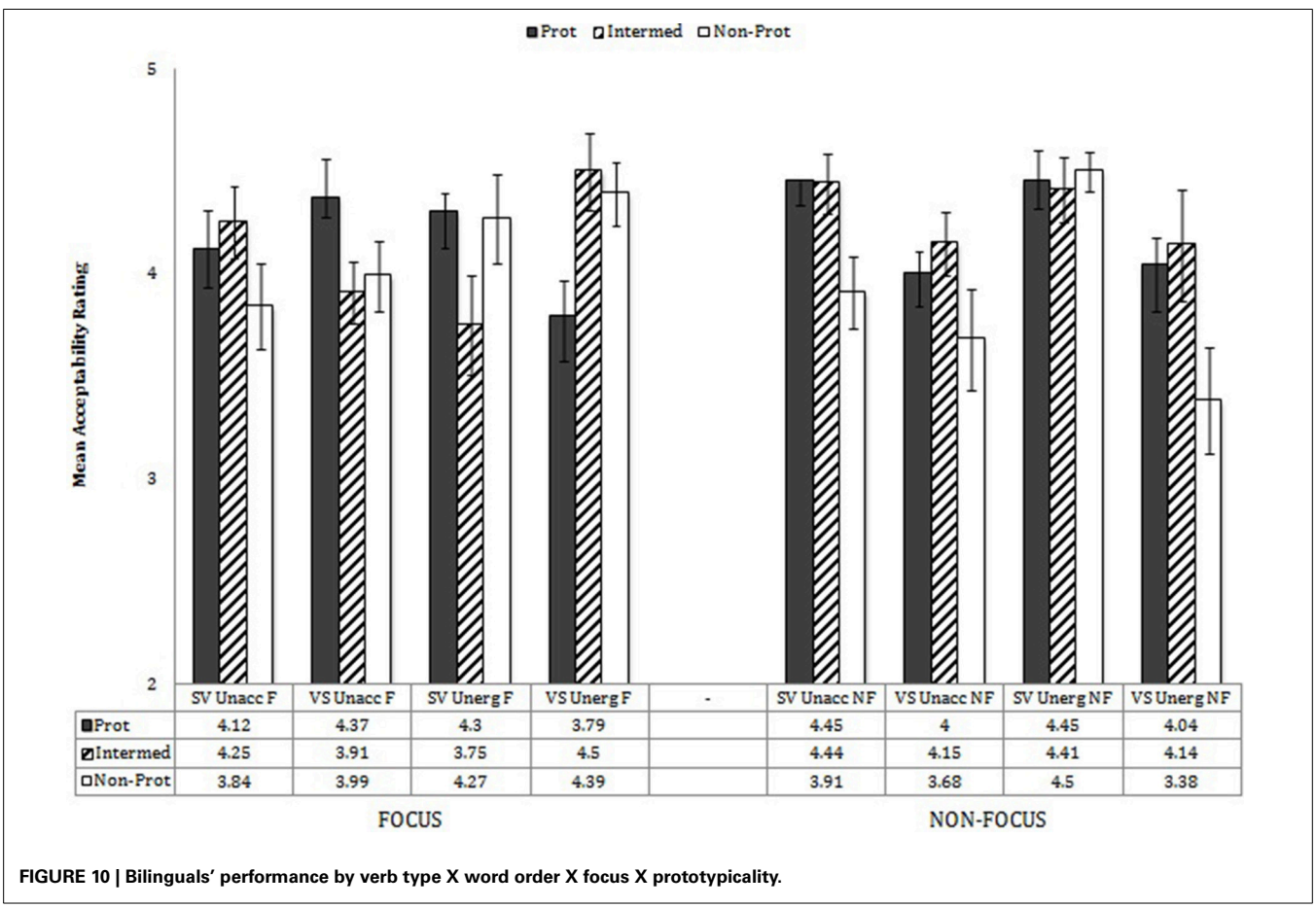


FIGURE 10 | Bilinguals' performance by verb type X word order X focus X prototypicality.

(1) The acceptability ratings for unaccusatives and unergatives were consistent with some of the predictions:

(a) In non-focus contexts, SV was the preferred order for all types of unergative verbs,
(b) In non-focus contexts, VS was judged slightly more acceptable with (prototypical and intermediate) unaccusative verbs than with (prototypical and intermediate) unergative verbs, and
(c) In focus contexts, VS order was preferred with (prototypical and non-prototypical) unaccusative verbs and with (intermediate and non-prototypical) unergative verbs.

(2) The acceptability ratings for unaccusatives and unergatives were not entirely as expected, however. In particular,

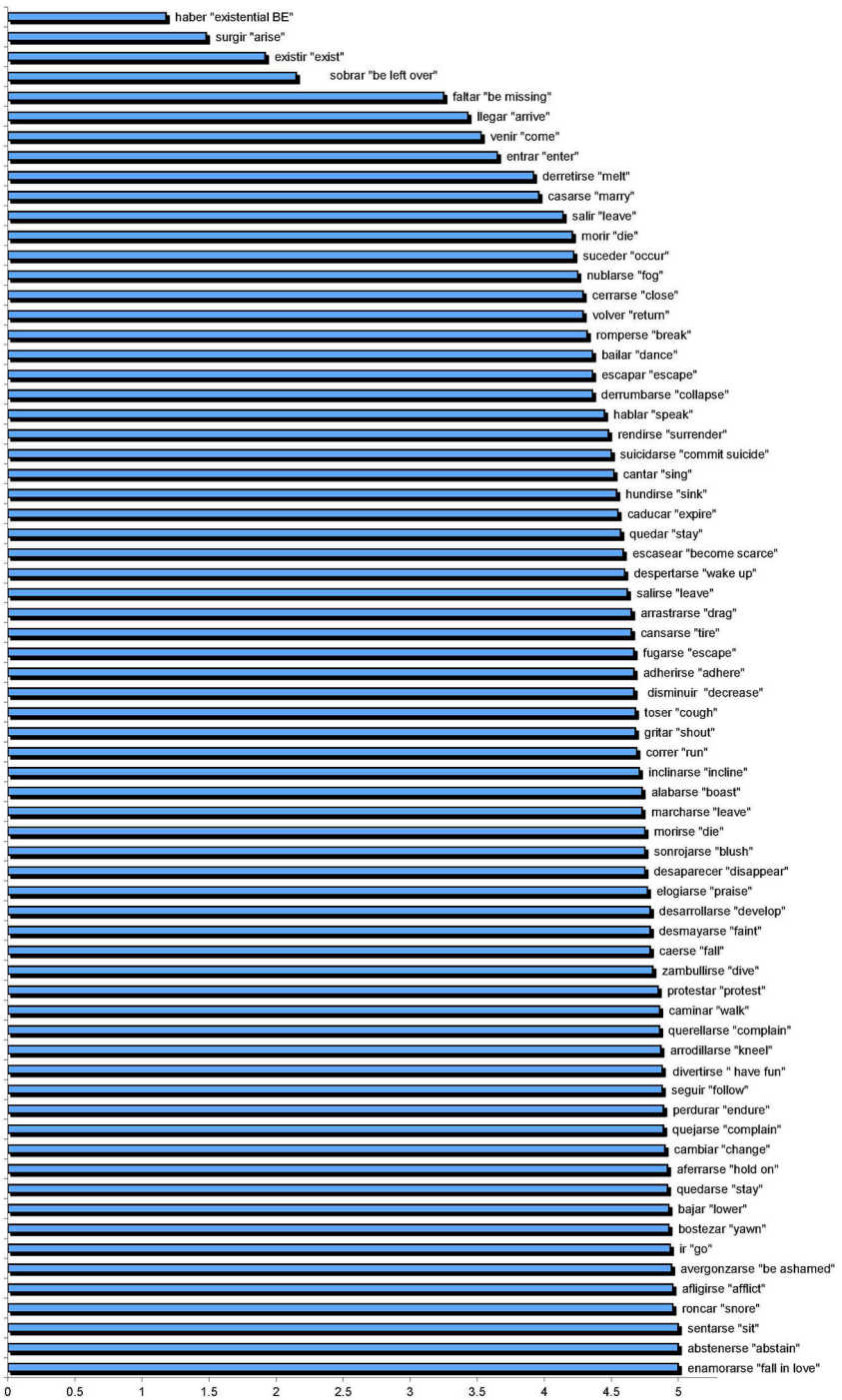


FIGURE 11 | Judgments, individual verbs in SV non-focus contexts.

(a) In non-focus contexts, SV order was the preferred order for both verb types, including unaccusatives,
(b) SV order was accepted for most verbs even in focus contexts (with the exception of non-prototypical unaccusative verbs and possibly intermediate unergative verbs),
(c) VS order was accepted in non-focus contexts more for prototypical and intermediate unergatives than for non-prototypical unergatives (similar to what was found for the unaccusatives), and
(d) VS order in focus contexts was treated as fairly unacceptable for prototypical unergative verbs.

With regard to the participant groups:

(3) First, the bilinguals did not distinguish unaccusative verbs from unergative verbs (unlike the monolinguals);
(4) The bilinguals did not distinguish SV from VS order in focus contexts (unlike the monolinguals); and
(5) The bilinguals' favoring of SV over VS order in non-focus contexts was less dramatic than the monolinguals' preference for SV in this context.
(6) Finally, at the same time, the bilinguals' performance relative to prototypicality levels of the particular verbs in particular contexts was similar on the whole to that of the monolinguals, except in two minor cases.

These results inform the questions at hand. First, with regard to the question of whether the bilinguals have come to a higher command of the verb type distinctions (internal interface) than of the operation of focus on syntactic structure (external interface), these data provide a resounding "no." Bilinguals did not on the whole differentiate the two verb types, and they only differentiated focus from non-focus contexts in that, whereas they accepted SV and VS orders equally in focus contexts, they favored SV order in non-focus contexts.

At the same time, however, it is clear that the bilinguals' performance was not random—their performance relative to particular verbs, as judged by the prototypicality effects, was similar to that of the monolinguals, but just at a less categorical or less extreme level.

Our findings with regard to the pervasiveness of use of SV with both unaccusative and unergative verbs, and the acceptance of VS even with unergative verbs, challenge the position that these verbs fall into a dichotomy. In order to explore these results further and to gain a better understanding of the findings, for each participant group, the performance on each verb was plotted and the verbs placed on a continuum.

The acceptability scores for each verb when it occurred with SV order in non-focus contexts are shown in **Figure 11** for the monolingual participants. Contrary to Lozano's claims that SV and VS do not alternate freely in native speaker grammars and that Spanish speakers treat the constructions categorically, our data show that there is a continuum, rather than a dichotomy.

The "true" unaccusatives should be those that were rated low when they occurred with SV order in non-focused contexts. The ten verbs with lowest acceptability ratings were those shown in **Figure 12**.

(Low SV in Non Focus Contexts).

The acceptability scores for each verb when it occurred with VS order in a non-focus context are shown in **Figure 13**.

It is interesting not only that the data reveal no clear-cut across-the-board binary distinction between unergatives and unaccusatives, but also that the data also do not follow Sorace's hierarchical semantic continuum. These results are in line with those reported in de Prada Pérez and Pascual y Cabo (2012). Such findings merit further exploration. What factors are influencing speakers' judgments of these constructs? One possibility has to do with exposure to the particular individual verbs in context. Perhaps speakers have more marked judgments in relation to verbs they experience more frequently. To examine this possibility, we extracted the frequency of occurrence of each verb

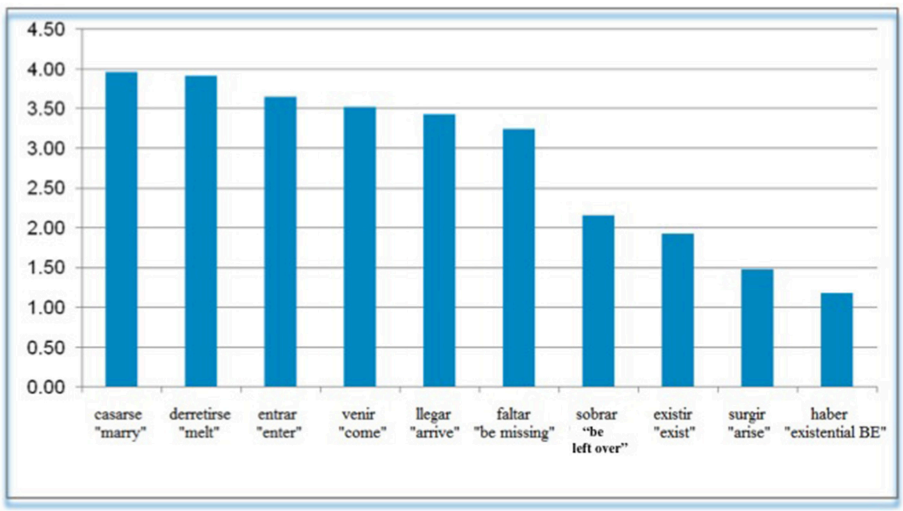


FIGURE 12 | True unaccusatives- lowest ratings for SV in non focus condition.

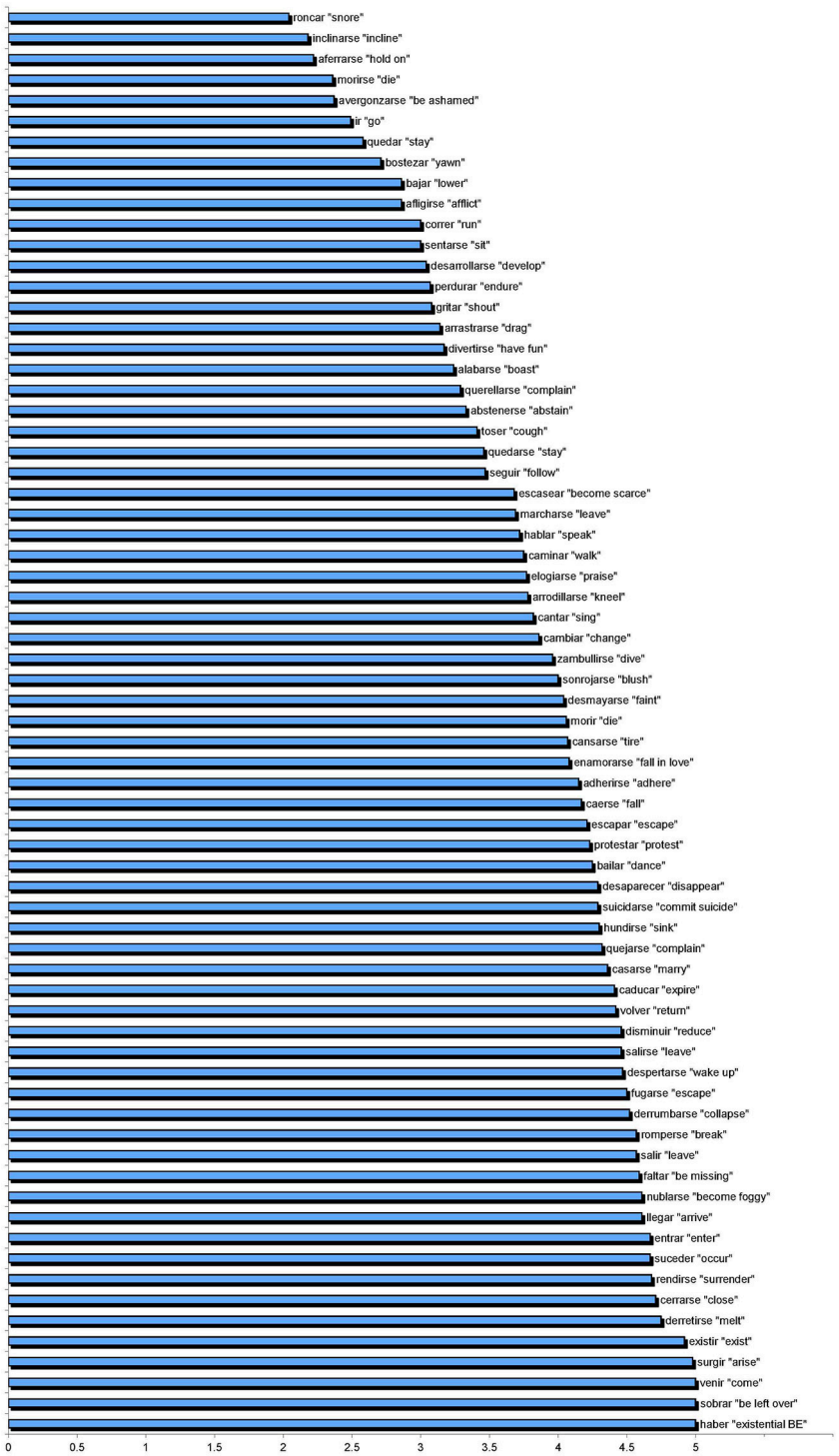


FIGURE 13 | Judgments, individual verbs in VS non-focus contexts.

(in all its forms) from SUBTLEX-ESP (Cuetos et al., 2011) (out of 41,577,673 words), and we conducted correlational analyses of these frequencies relative to performance in each of the major contexts—SV Focus (SVF), SV non-Focus (SVnF), VS Focus (VSF), and VS non-Focus (VSnF). These analyses were conducted, first, with verbs of all types together, and then with the unergative and unaccusative verbs separately, for both the Monolinguals and the Bilinguals.

For the Monolinguals, first, for all verbs together ($N = 69$), the speakers' judgments showed a high correlation between

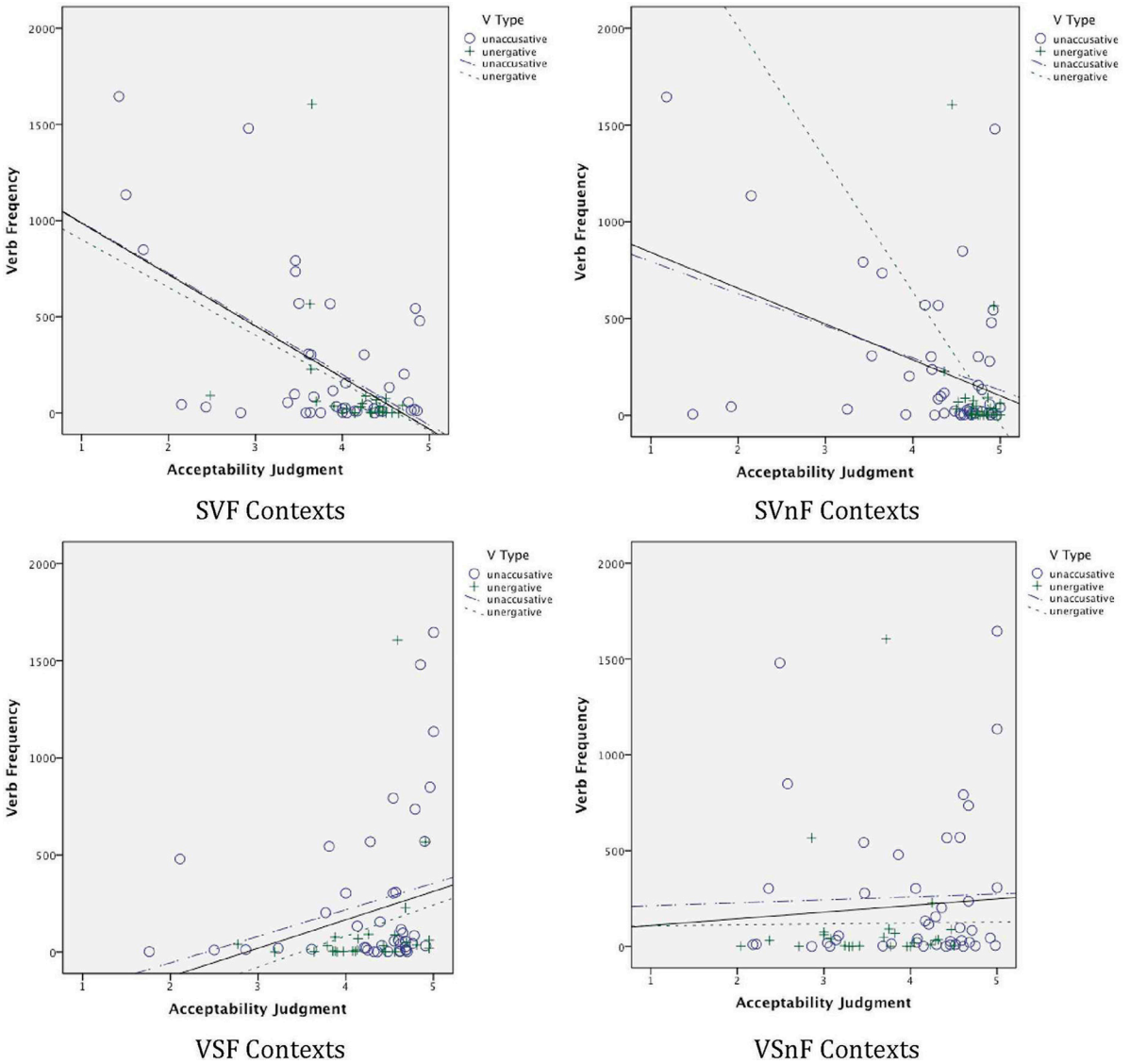


FIGURE 14 | Monolingual judgments.

judgments in the SVF, SVnF, and VSF contexts with the frequency of the verb. In SVF and SVnF settings, the correlation was negative, $r = -0.544$, $p < 0.001$, and $r = -0.384$, $p = 0.001$, respectively, indicating that the more frequent the verb, the less they accepted the SV order in both F and nF contexts. In VSF contexts, the correlation was positive, $r = 0.260$, $p = 0.031$, indicating that the more frequent the verb, the greater the acceptance of the VS order in F contexts. For unaccusative verbs alone ($N = 43$), negative correlations still held for SVF and SVnF, at $r = -0.577$, $p < 0.001$, and $r = -0.385$, $p = 0.010$, respectively, but the positive correlation in relation to VSF did not reach significance ($r = 0.254$, $p = 0.100$). For unergative verbs alone ($N = 26$), none of these reached significance, although for SVF and SVnF contexts, the correlations were near-significant, $r = -0.354$, $p = 0.076$, and $r = -0.362$, $p = 0.076$, respectively. Scatter plots showing the Monolinguals' judgments relative to verb frequency in each type of context are shown in **Figure 14**. (The slopes for all verbs together are shown with solid lines; those for the unaccusatives and unergatives separately with dotted lines as indicated.)

For the Bilinguals, for all verbs together ($N = 72$), there was a negative correlation between judgments in SVF contexts and frequency, $r = -0.318$, $p = 0.006$, again indicating that the more frequent the verb, the lower the judgments were for the verb in SV order in F contexts. There was also a positive correlation of frequency of the verb with acceptance of VS order in nF contexts, $r = 0.260$, $p = 0.027$. This latter result is surprising, as it indicates that the frequency with which a verb is heard correlates with higher acceptance of VS order even in non-Focus contexts. Neither of these correlations held for unergative verbs alone ($N = 26$), but for unaccusative verbs ($N = 46$), the negative correlation held in relation to SVF contexts, $r = -0.320$, $p = 0.030$. Scatter

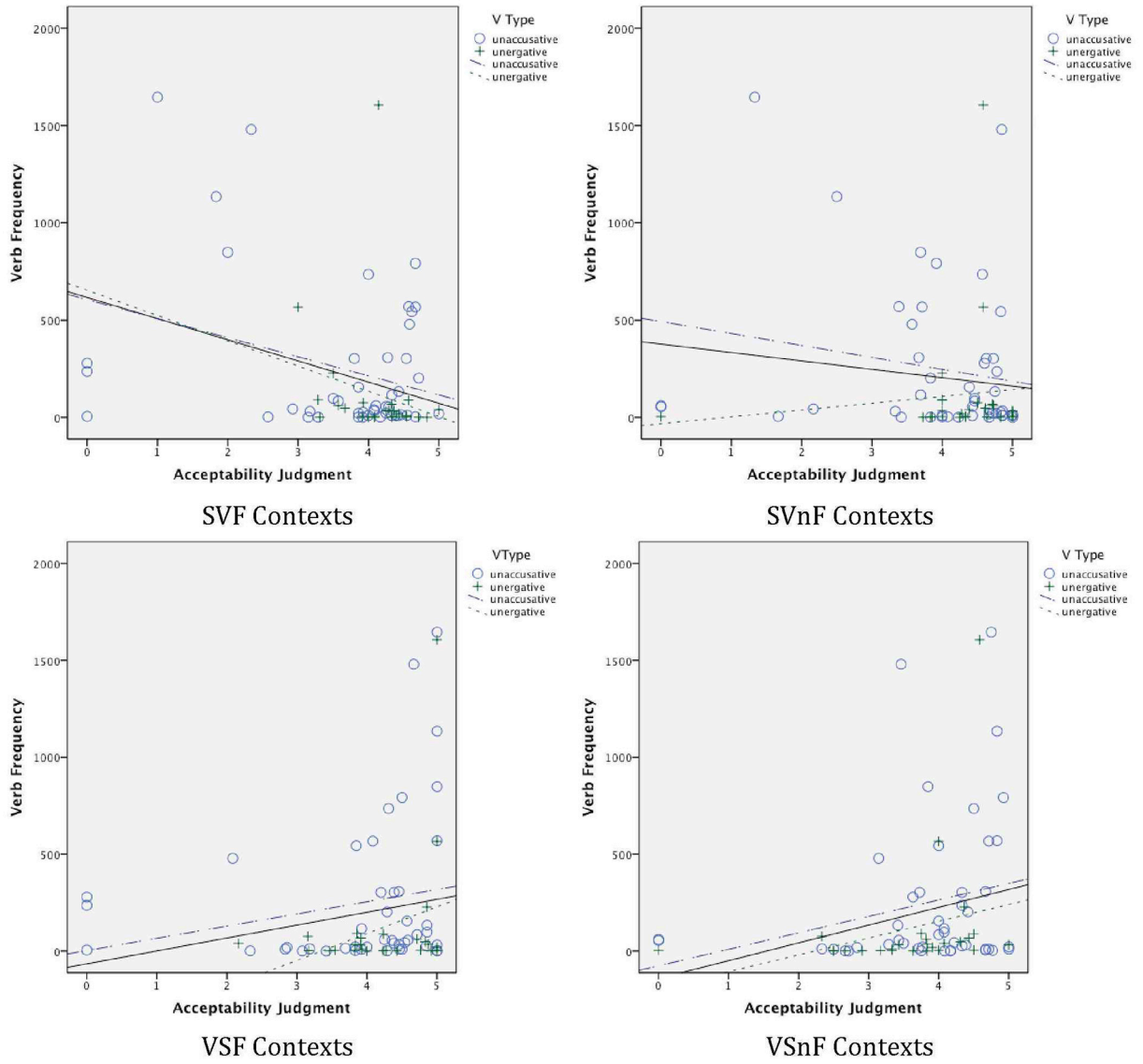


FIGURE 15 | Bilingual judgments.

plots of the Bilinguals' judgments relative to verb frequency are shown for each context in **Figure 15**.

These correlations suggest that judgments are highly influenced by exposure to the particular verb in the given construction. For the Monolinguals, the more frequent the verb, the less the acceptance of SV order, in both F and nF contexts (the latter especially in the case of unaccusative verbs). Similarly, the more frequent the verb, the more they accept VS order in F contexts. The Bilinguals show a similar effect in SVF contexts—the more frequent the verb, the less they accept this order in F contexts. The fact that the Bilinguals pattern like the Monolinguals in these SVF contexts is further support for the conclusion drawn above that Bilinguals have in fact acquired features of the grammar related to the interaction of the discourse and syntax. Furthermore, all of these effects are in line with grammatical accounts regarding SV and VS order in F contexts—with speakers disfavoring SV order in F contexts for presumably well-ingrained verbs.

The similarity of verb patterns for unaccusative and unergative verbs, however, argues against a strong dichotomy between the two verb types, in line with what we have argued above. The effect showing that the Bilinguals are more likely to be more accepting of VSnF structures for more frequent verbs is difficult to explain, however. The effect is weak, but is deserving, as are the other correlations discovered, of more targeted research in future work.

CONCLUSIONS

This study has explored syntactic, pragmatic, and lexical influences on adherence to SV and VS orders in native and fluent L2 speakers of Spanish who were long-standing functional bilinguals. The primary issue addressed has been the hypothesis that

bilinguals have no difficulty in acquiring within-module grammatical elements and only encounter difficulties in relation to external interface phenomena. The data presented here do not support this position. Here, first, the bilinguals treated unaccusative and unergative verbs in identical fashion; this indicates that they have not gained a native-like command of this feature. In contrast, long-term L2 bilinguals did differentiate between focus and non-focus, in that they preferred SV order in non-focus contexts. Hence, our findings do not support the hypothesis that internal interfaces are acquired more easily than external interfaces (contra accepted wisdom- Sorace, 2005; Sorace and Filiaci, 2006; White, 2006).

A further result of this study concerns the classification of verbs themselves. Unlike previous empirical studies (Hertel, 2003; Lozano, 2006a,b) and contra the theoretical literature (Contreras, 1978; Suñer, 1982 and Zubizarreta, 1998), the data revealed no clear-cut across-the-board binary distinction between unergatives and unaccusatives (but see Domínguez and Arche, 2014). Neither monolinguals nor bilinguals differentiated the two types of verbs in non-focus situations. However, monolinguals paid attention to verb type in the focus situation in that they preferred VS for unaccusatives. Long-term L2 bilinguals did not differentiate between the two types of verbs, treating unaccusatives and unergatives equally even in focus contexts. The follow-up analyses showed strong correlations between speakers' judgment performance with the frequency of the verbs. These findings suggest that any account of the use of these verbs in these contexts will need to take into account a usage-based perspective. That is, it appears that performance on each verb is determined at least in part by speakers' relative experience with the verbs in question, rather than strict verb categories based either on syntax or semantics.

Two additional findings of the study are worth noting. First, bilinguals' judgments were less categorical overall than monolinguals'. That is, monolinguals were more likely to give extreme "yes" or "no" judgments, while bilinguals gave more intermediate judgments (cf. lower confidence ratings by bilinguals in their judgments compared to monolinguals in Sagarra and Herschensohn (2013). Second, individual verbs do not necessarily behave as predicted under standard definitions of unaccusatives and unergatives.

The data presented here do not lend support to a split intransitivity dichotomy. Rather, they support a continuum. This continuum, however, does not seem to fit Sorace's criteria defined primarily by aspectual notions (telicity/atelicity), and secondarily by the degree of agentivity of the verb. The functional bilinguals' performance differed from the performance of monolinguals in their lack of differentiation of verb types. And contrary to what is predicted according to the interface hypothesis, bilinguals differentiated between focus and non-focus situations.

Our results challenge the second version of the interface hypothesis (Sorace, 2011) by postulating that external interfaces (syntax-discourse) are not necessarily more difficult than internal interfaces (lexicon-syntax). They also provide new empirical evidence on the Unaccusative Hierarchy (semantic subclasses of unaccusatives and unergatives) for native Spanish, which does not work in the same way as proposed for Italian. In addition, the results of this study shed new light on word order alternations (SV/VS) which, although previously studied, needed to be investigated in more detail in long-standing functional bilinguals if we wanted a better understanding of this phenomenon in end-state grammars.

If confirmed and broadened through further research using a wider range of methodologies with learners at different levels of proficiency, the findings of this study have fundamental implications for our understanding of the interface system in L2 learners and for our general understanding of the grammar of unaccusative and unergative verbs.

ACKNOWLEDGMENTS

We acknowledge the support of an ESRC Centre for Research on Bilingualism, from which the authors received a Development Fund to conduct this research. We thank Lexi Hindley, Noriko Hoshino, Diego Muñoz-Carrobles, Miguel Ángel Pérez, Rocío Pérez Tattam, Eugenia Sebastián, Carmen Silva-Corvalán, Pilar Soto, and Patricia Tattam for help with data collection and general discussion.

REFERENCES

Baker, C. (1993). *Foundations of Bilingual Education and Bilingualism*. Clevedon: Multilingual Matters.

Baker, M. (1983). "Objects, themes and lexical rules in Italian," in *Papers in Lexical-Functional Grammar,*" eds L. Levin, M. Rappaport Hovav, and A. Zaenen (Bloomington, IN: Indiana University Linguistics Club), 1–46.

Beck, M. L. (1998). L2 acquisition and obligatory head movement: English-speaking learners of German and the local impairment hypothesis. *Stud. Second Lang. Acquis.* 20, 311–348.

Belletti, A. (2001). "'Inversion' as focalization," in *Inversion in Romance and the Theory of Universal Grammar,* eds A. Hulk and J. Y. Pollock (Oxford: Oxford University Press), 60–90.

Belletti, A. (2004). "Aspects of the low IP area," in *The Structure of IP and CP. The cartography of Syntactic Structures,* Vol. 2, ed L. Rizzi (Oxford: Oxford University Press), 16–51.

Belletti, A., and Leonini, C. (2004). "Subject inversion in L2 Italian," in *Eurosla Yearbook 4*, eds S. Foster-Cohen, M. Sharwood Smith, A. Sorace, and M. Ota (Amsterdam: John Benjamins Publishing Company), 95–118.

Burzio, L. (1986). *Italian Syntax: A Government and Binding Approach*. Dordrecht: Reidel.

Camacho, J. (1999). From SOV to SVO: the grammar of interlanguage word order. Second *Lang. Res.* 15, 115–132.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: The MIT Press.

Chomsky, N. (2005). Three factors in language design. *Linguist. Inq.* 36, 1–22. doi: 10.1162/0024389052993655

Contreras, H. (1978). *El orden de Palabras en Español*. Madrid: Cátedra.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: SAGE.

Cuetos, F., Glez-Nosti, M., Barbon, A., and Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica* 32, 133–143.

De Miguel, E. (1993). "Construcciones ergativas e inversión en la lengua y la interlengua españolas," in *La lingüística y el Análisis de Los Sistemas No Nativos,* ed J. Liceras (Ottawa, ON: Dovehouse), 178–195.

de Prada Pérez, A., and Pascual y Cabo, D. (2012). "Interface heritage speech across proficiencies: unaccusativity, focus, and subject position in Spanish," in *Selected Proceedings of the 14th Hispanic Linguistics Symposium,* eds K. Geeslin and M. Díaz-Campos (Somerville, MA: Cascadilla Proceedings Project), 308–318.

Domínguez, L., and Arche, M. J. (2014). Subject inversion in non-native Spanish. *Lingua* 145, 243–265. doi: 10.1016/j.lingua.2014.04.004

Dowty, D. R. (1991). Thematic proto-roles and argument selection. *Language* 67, 547–619.

Friedmann, N., Taranto, G., Shapiro, L. P., and Swinney, D. (2008). The vase fell (the vase): the online processing of unaccusatives. *Linguist. Inq.* 39, 355–377. doi: 10.1162/ling.2008.39.3.355

Hatcher, A. (1956). Theme and underlying question: two studies of Spanish word order. *Word* 12, 3.

Hawkins, R. (2000). Persistent selective fossilization in second language acquisition and the optimal design of the language faculty. *Essex Res. Rep. Linguist.* 34. 75–90.

Hertel, T. J. (2003). Lexical and discourse factors in the second language acquisition of Spanish word order. Second *Lang. Res.* 19, 273–304. doi: 10.1191/0267658303sr224oa

Levin, B., and Rappaport Hovav, M. (1995). *Unaccusativity at the Syntax-Lexical Semantics Interface*. Cambridge, MA: MIT Press.

Lozano, C. (2003). *Universal Grammar and Focus Constraints: The Acquisition of Pronouns and Word Order in Non-Native Spanish*. PhD dissertation. University of Essex.

Lozano, C. (2006a). Focus and split intransitivity: the acquisition of word order alternations in non-native Spanish. *Second Lang. Res.* 22, 1–43. doi: 10.1191/0267658306sr264oa

Lozano, C. (2006b). "The development of the syntax-discourse interface: Greek learners of Spanish," in *The Acquisition of Syntax in Romance Languages,* eds V. Torrens and L. Escobar (Amsterdam: John Benjamins), 371–399.

Lozano, C. (2009). *The Acquisition of Syntax and Discourse: Pronominals and Word Order in English and Greek Learners of Non-Native Spanish*. Saarbrüken: VDM Verlag.

Montrul, S. (2001). Causatives and transitivity in L2 English. *Lang. Learn.* 51, 51–106. doi: 10.1111/1467-9922.00148

Montrul, S. (2004). Subject and object expression in Spanish heritage speakers: a case of morphosyntactic convergence. *Bilingualism Lang. Cogn.* 7, 125–142. doi: 10.1017/S1366728904001464

Montrul, S. (2005). On knowledge and development of unaccusativity in Spanish L2 acquisition. *Linguistics* 43, 1153–1190. doi: 10.1515/ling.2005.43.6.1153

Nava, E. (2007). "Word order in bilingual Spanish: convergence and intonation strategy," in *Selected Proceedings of the Third Workshop on Spanish Sociolinguistics,* eds J. Holmquist, A. Lorenzino, and L. Sayahi (Somerville, MA: Cascadilla Proceedings Project), 129–139.

Ocampo, F. (1990). The pragmatics of word order in constructions with a verb and a subject. *Hisp. Linguist.* 4, 87–127.

Oshita, H. (2001). The unaccusative trap in second language acquisition. *Stud. Second Lang. Acquis.* 23, 279–304. doi: 10.1017/S0272263101002078

Perlmutter, D. M. (1978). "Impersonal passives and the unaccusativity hypothesis," in *Proceedings of the Fourth Annual Meeting of the Berkeley Linguistic Society* (Berkeley: Berkeley Linguistic Society; University of California), 157–189.

Perlmutter, D. M., and Postal, P. M. (1984). "The I-advancement exclusiveness law," in *Studies in Relational Grammar 2,* eds D. M. Perlmutter and C. Rosen (Chicago, IL: University of Chicago Press), 81–125.

Pinto, M. (1999). "Information focus: between core and periphery," in *Semantic Issues in Romance Syntax,* eds E. Treviño and J. Lema (Amsterdam: John Benjamins), 179–191.

Rappaport Hovav, M., and Levin, B. (2001). An event structure account of English resultatives. *Language* 77, 766–797. doi: 10.1353/lan.2001.0221

Rosen, C. (1984). "The interface between semantic roles and initial grammatical relations," in *Studies in Relational Grammar* eds D. Perlmutter and C. Rosen (Chicago, IL: University of Chicago Press), 38–77.

Sagarra, N., and Herschensohn, J. (2013). Processing of gender and number agreement in late Spanish bilinguals. *Int. J. Bilingualism* 17, 607–627. doi: 10.1177/1367006912453810

Schmerling, S. (1976). *Aspects of English Sentence Stress.* Austin, TX: University of Texas Press.

Selkirk, E. (1984). *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, MA: MIT Press.

Sorace, A. (1993a). Incomplete vs. divergent representations of unaccusativity in non-native grammars of Italian. *Second Lang. Res.* 9, 22–47.

Sorace, A. (1993b). Unaccusativity and auxiliary choice in non-native grammars of Italian and French: asymmetries and predicable indeterminacy. *J. Fr. Lang. Stud.* 3, 71–93.

Sorace, A. (1995). "Acquiring linking rules and argument structures in a second language: the unaccusative/unergative distinction," in *The Current State of Interlanguage,* eds L. Eubank, L. Selinker, and M. S. Sharwood (Amsterdam: Benjamins), 153–175.

Sorace, A. (2000). Gradients in auxiliary selection with intransitive verbs. *Language* 76, 859–890. doi: 10.2307/417202

Sorace, A. (2005). "Syntactic optionality at interfaces," in *Syntax and Variation: Reconciling the Biological and the Social,* eds L. Cornips and K. Corrigan (Amsterdam: John Benjamins), 46–111.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguist. Approaches Bilingualism* 1, 1–33. doi: 10.1075/lab.1.1.01sor

Sorace, A., and Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. Second *Lang. Res.* 22, 339–368. doi: 10.1191/0267658306sr271oa

Sorace, A., and Shomura, Y. (2001). Lexical constraints on the acquisition of split intransitivity: evidence from L2 Japanese. *Stud. Second Lang. Acquis.* 23, 247–278. doi: 10.1017/S0272263101002066

Suñer, M. (1982). *Syntax and Semantics of Spanish Presentational Sentence-types*. Washington, DC.: Georgetown University Press.

Tsimpli, I., and Sorace, A. (2006). "Differentiating interfaces: L2 performance in syntax-semantics and syntax-discourse phenomena," in *Proceedings of the 30th Annual Boston University Conference on Language Development* (Somerville, MA: Cascadilla Press), 653–664.

van Hout, A. (1996). *Event Semantics of Verb Frame Alternations: A Case Study of Dutch and its Acquisition*. Doctoral dissertation, Tilburg University.

van Hout, A., Randall, J., and Weissenborn, J. (1992). "Acquiring the unaccusative/unergative distinction," in *The Acquisition of Dutch Amsterdam Series in Child Language Development 1,* eds M. Verrips and F. Wijnen (Amsterdam: University of Amsterdam), 79–120.

van Valin, R. (1999). "Generalized semantic roles and the syntax-semantics interface," in *Empirical Issues in Formal Syntax and Semantics 2,* eds F. Corblin, C. Dobrovie-Sorin and J.-M. Marandin (The Hague: Thesus), 373–389.

White, L. (2006). "Interfaces and L2 knowledge: the Spanish connection,"in *Plenary address presented at the Hispanic and Luso-Brazilian Linguistics Symposium* (London: The University of Western Ontario).

White, L. (2009). "Language acquisition at the interfaces: some hardy perennials and new varieties," in *Mind-Context Divide Conference* (Iowa City: University of Iowa).

Yip, V. (1995). *Interlanguage and Learnability: From Chinese to English*. Amsterdam: Benjamins.

Zubizarreta, M. L. (1998). *Prosody, Focus, and Word Order*. Cambridge, MA: MIT Press.

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: 09 October 2014; accepted: 10 December 2014; published online: 13 January 2015.

*Citation: Parafita Couto MC, Mueller Gathercole VC and Stadthagen-González H (2015) Interface strategies in monolingual and end-state L2 Spanish grammars are not that different. Front. Psychol. **5**:1525. doi: 10.3389/fpsyg.2014.01525*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 13 May 2014
doi: 10.3389/fpsyg.2014.00424

Discriminating languages in bilingual contexts: the impact of orthographic markedness

Aina Casaponsa[1]*, Manuel Carreiras[1,2,3] and Jon A. Duñabeitia[1]

[1] *Basque Center on Cognition, Brain and Language, Donostia, Spain*
[2] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
[3] *University of the Basque Country EHU/UPV, Bilbao, Spain*

Edited by:
Christos Pliatsikas, University of Kent, UK

Reviewed by:
Jonathan Grainger, CNRS, France
Davide Crepaldi, University of Milano-Bicocca, Italy

****Correspondence:***
Aina Casaponsa, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd floor, 20009 Donostia, Spain
e-mail: a.casaponsa@bcbl.eu

Does language-specific orthography help language detection and lexical access in naturalistic bilingual contexts? This study investigates how L2 orthotactic properties influence bilingual language detection in bilingual societies and the extent to which it modulates lexical access and single word processing. Language specificity of naturalistically learnt L2 words was manipulated by including bigram combinations that could be either L2 language-specific or common in the two languages known by bilinguals. A group of balanced bilinguals and a group of highly proficient but unbalanced bilinguals who grew up in a bilingual society were tested, together with a group of monolinguals (for control purposes). All the participants completed a speeded language detection task and a progressive demasking task. Results showed that the use of the information of orthotactic rules across languages depends on the task demands at hand, and on participants' proficiency in the second language. The influence of language orthotactic rules during language detection, lexical access and word identification are discussed according to the most prominent models of bilingual word recognition.

Keywords: bilingual reading, visual word recognition, orthographic cues, bigrams, selective lexical access

INTRODUCTION

Bilingual societies in which two official languages coexist (e.g., Basque Country, Catalonia, Wales) have attracted a great deal of scientific attention in recent decades, given that balanced simultaneous bilinguals who are exposed to two languages on a daily basis can provide evidence about the organization of bilingual lexical representations and the mechanisms leading to effective language selection in naturalistic contexts (e.g., Costa et al., 2006; Perea et al., 2008; Duñabeitia et al., 2010b; Kuipers and Thierry, 2010). So far the evidence regarding the differences and similarities between the mechanisms guiding lexico-semantic and syntactic processing in non-simultaneous bilinguals who learn the L2 in naturalistic vs. classroom contexts is still mixed (see Muñoz, 2008; Pliatsikas and Marinis, 2013, for reviews). However, results from studies testing early simultaneous balanced bilinguals living in bilingual contexts offer converging evidence on the effectiveness of cross-language activation at multiple levels of processing, especially in the visual word recognition domain (e.g., Thierry and Wu, 2007; Duñabeitia et al., 2010a). Although there is evidence demonstrating that the recognition of a visually presented word is governed by parallel access to both languages used by balanced simultaneous bilinguals, the mechanisms by which a given word form is associated with a given language (i.e., language tagging) by these bilinguals is still unclear.

In most bilingual environments readers can find different cues that help bilingual language recognition and lexical access. One extreme example of this reality is the case of languages that do not share the same script (e.g., Hebrew-English), since the individual letters that constitute the printed words are the clearest language cue. However, this situation does not apply to multiple bilingual societies in which both languages are highly similar and share the same orthography (e.g., French-English, Spanish-Basque), therefore making it difficult for readers to determine the language of each individual word. Hence, studying bilingual visual word recognition with same-script language combinations may help us to identify which are the features of the words that aid bilingual language selection and recognition.

Different languages have different orthotactic rules, and it seems plausible to assume that bilinguals could rely in such cues as a strategy while reading in an ambiguous language context. In fact, previous studies have suggested that the frequency of the letters and their combinations within a language may play an important role in bilingual language detection and to some extent may also mediate the lexical access process (Grainger and Beauvillain, 1987; Thomas and Allport, 2000; Vaid and Frenck-Mestre, 2002; Lemhöfer et al., 2008, 2011; Van Kesteren et al., 2012). Vaid and Frenck-Mestre (2002) presented English and French words to highly proficient English-French bilinguals in a speeded language decision task, and found that words that were clearly marked as belonging to one of the languages in terms of bigram frequencies (e.g., OEUF as a French-marked word) were responded to faster than unmarked words (words that follow the same orthotactic rules in both languages). These results suggest that language decision or detection could be mediated by the extraction of statistical orthographic regularities at early stages of single word processing.

In a similar vein, a recent study by Lemhöfer et al. (2011) testing compound words in a lexical decision task with monolinguals

and bilinguals showed clear-cut orthographic markedness effects. They presented native and non-native Dutch participants with Dutch compound words that could contain an orthotactic parsing cue (i.e., the bigram at the morphemic boundary being a bigram that cannot exist within a Dutch morpheme), and found that the presence of such parsing cue aided morphological decomposition. In line with the results presented by Vaid and Frenck-Mestre (2002), Lemhöfer et al. concluded that the sub-lexical information at the constituent boundary might guide the identification of the individual constituents, thus helping word recognition.

Recently, Van Kesteren et al. (2012) demonstrated a language decision advantage for words that contain language-specific orthography in one of the bilingual languages, proposing a direct link between sub-lexical information of words and language membership. In their study, Norwegian-English bilinguals completed a series of language decision and lexical decision tasks including marked and unmarked Norwegian and English words, and their results demonstrated a strong reliance of bilingual readers on sub-lexical orthographic properties of words, given the clear-cut markedness effects found across tasks. They concluded that language information could be accessed directly via sub-lexical information instead of via lexical representation of words, and they proposed an extension of the Bilingual Interactive Activation Plus model in order to account for these effects. These findings closely match earlier evidence demonstrating that language-specific orthography directly affects single word identification (e.g., Vaid and Frenck-Mestre, 2002; Lemhöfer et al., 2008, 2011), suggesting that access to the lexicon might be guided by the extraction of language-specific orthotactic combinatorial rules. That is, at early stages of visual word identification, language selection mechanisms have been proposed to operate enhancing lexical activation of the relevant language on the basis of the sub-lexical structure of the words (see Grainger and Beauvillain, 1987; Schwartz et al., 2007; see also Westbury and Buchanan, 2002, for evidence in monolinguals).

The main goal of the current study is to better understand bilingual language identification processes by exploring the influence exerted by the sub-lexical characteristics of the words in bilinguals' two languages during different stages of word recognition. We tested L2 words that could be either legal or illegal in the L1 vocabulary in terms of their corresponding bigram frequencies in two different tasks with varying demands of lexical access: a language decision task and a perceptual identification task. To this end, the materials included Basque-specific words (e.g., ETXE [house]; note that TX is an illegal bigram in Spanish) and orthographically unmarked words (e.g., MENDI [hill]; note that all the bigrams are also plausible in Spanish). Besides, in order to explore whether or not the reliance on L2 orthotactic cues depends on the degree of L2 proficiency, three different samples of participants were tested: balanced Spanish-Basque bilinguals, highly-proficient unbalanced bilinguals (L1 Spanish, L2 Basque) and Spanish monolinguals.

Interestingly, and contrary to any explanation of the markedness effect in terms of sub-lexical-lexical interactions, Vaid and Frenck-Mestre (2000) proposed that the markedness effect "*is predominantly a perceptual effect rather than one involving complete lexical access*" (p. 52). It should be noted that the term "perceptual" refers to an effect based on orthographic information of words rather than on perceptual features of letters. That is, they suggested that participants could have taken their language decisions following a sub-lexical strategy, without relying on the real meaning of the marked words. Rather than discriminating L1 and L2 based on the lexical representations in each language, participants could have taken a different strategy and could have completed the task by simply deciding whether or not a given string corresponds to the L1. In other words, rather than identifying OUEF as a French word by accessing its meaning, participants could have simply discarded it as an English word given its orthotactic regularities. Such an account is difficult to reject on the basis of the existing evidence. However, one possible solution to the conundrum is to include a group of monolinguals in the experiment and to compare their performance to that of bilinguals. If participants exclusively rely on low-level L1 orthographic rules to perform language discrimination tasks, even monolinguals would benefit of the presence of L2-marked words, showing facilitative effects for strings containing orthotactic cues regardless of their L2 knowledge. Put differently, if the L2 markedness effect exclusively relies on a sub-lexical strategy based on the detection of L1 orthographic violations, extensive knowledge of the L2 is not required to complete the language detection task, given the sub-lexical locus of the decision criteria. Hence, even monolinguals who are not familiar with the L2 could perform correctly on the basis of this account. At the same time an inhibitory effect is expected for monolinguals compared to bilinguals for non-marked L2 words due to the similarity with real words (see Westbury and Buchanan, 2002; Lemhöfer et al., 2008, for a review).

As previously mentioned, most of the studies exploring the L2 word markedness effect have used tasks that do not explicitly require full lexical access to the written representations, given that these studies have mainly used the language decision task or the (mixed) language lexical decision task (see Vaid and Frenck-Mestre, 2002; Van Kesteren et al., 2012). One of the main problems with these tasks is that it is difficult to estimate the degree of lexical access needed to efficiently determine whether a given string corresponds to language X or Y, or whether it is a real or invented word, given the difficulty to estimate the impact of factors associated with word likelihood (e.g., Jacobs and Grainger, 1994; Jacobs et al., 1998; see Wagenmakers et al., 2004, for review). Hence, in order to disambiguate between proposals claiming for a different influence of L2 orthotactic cues in sub-lexical ortgographic decisions, on the one hand, and in lexico-semantic access, on the other, and following the line opened by Van Kesteren et al. who suggested that the L2 markedness effect may largely depend on the specific task demands, in the present study we investigated the presence of this effect in two tasks, one of which explicitly requires conscious access to the specific visually presented representation. Participants' performance in a language decision task (Experiment 1) was compared to their performance in a perceptual identification task (Experiment 2). The perceptual identification task selected was the progressive demasking task (PDM hereafter) developed by Grainger and Segui (1990) and implemented by Dufau et al. (2008). The PDM is a perceptual task that requires participants to recognize letter strings

by pressing a button key and to write them back on the keyboard (for different applications of this task, see Carreiras et al., 1997; Duñabeitia et al., 2008). Importantly, the PDM task does not allow for responses based on a mere strategy of estimating the L1-membership likelihood on the basis of specific letter combinations, since the whole string needs to be retained in memory to correctly complete the task. Given the difficulty to access and remember L2-marked strings for monolingual participants who presumably have never faced the critical L1-illegal bigrams, it seems reasonable to tentatively predict that marked words would help bilinguals' performance in this task, while the opposite pattern is expected for monolingual participants. Besides, since monolingual participants will need to complete this task by following an orthography-to-phonological working memory strategy instead of a lexical strategy, they would take longer to recognize letter strings including letter combinations that are not present in their L1 as compared to words that follow the L1 orthographic rules (i.e., L2 unmarked words). It was also expected that, overall, L2 words would be harder to recognize for unbalanced than for balanced bilinguals, given that the speed and accuracy of lexical access is highly sensitive to proficiency (see, among many others, Dimitropoulou et al., 2011; Francis et al., 2014).

EXPERIMENT 1: SPEEDED LANGUAGE DECISION TASK

MATERIALS AND METHODS

Participants

Sixty undergraduates (44 women; mean age = 23.11, *SD* = 3.70) with normal or corrected-to-normal vision participated in this experiment in exchange for monetary compensation. Twenty were balanced Spanish-Basque bilinguals from the Basque Country (12 women; mean age = 24.54, *SD* = 5.29). These balanced bilinguals had a native-like proficiency in both Basque and Spanish, as calculated by their proficiency self-ratings (see **Table 1**). A group of 20 unbalanced bilinguals was also selected, being all of them native Spanish speakers from the Basque Country (14 women; mean age = 21.73, *SD* = 2.78) who learnt Basque as a second language and were relatively high proficient in Basque, but not native-like (see **Table 1**). The remaining 20 participants were Spanish monolinguals (18 women; mean age = 23.05, *SD* = 3.03) with no prior knowledge of Basque. The overall self-perception level of Spanish ranged from 9 to 10 for all groups of participants (mean = 9.73, *SD* = 0.45). Balanced bilinguals also ranged from 9 to 10 in their knowledge of Basque (mean = 9.62, *SD* = 0.50), and unbalanced bilinguals ranged from 6 to 8 in their self-perceived Basque proficiency (mean = 7.54, *SD* = 0.71). The monolinguals had never learnt Basque and all of them lived in Murcia, a monolingual region of Spain. None of the participants reported neurological or psychiatric disorders. All participants gave their written informed consent in accordance with guidelines approved by the Ethics and Research Committees of the Basque Center on Cognition, Brain and Language. The study was also performed in accordance with the ethical standards set in the Declaration of Helsinki.

Stimuli

Six hundred and eighty words were used as targets. Half of them were Spanish words taken from Davis and Perea (2005) and the other half were Basque words taken from Perea et al. (2006). Critically, Basque words were selected as a function of their bigram combinations so that they could be either valid o invalid in both languages. Half of the Basque words were marked by bigram combinations that were only plausible in Basque (i.e., L2-marked words; e.g., *txakur* [dog], where the bigram "tx" do not exist in Spanish), and the other half were unmarked words that also followed the Spanish orthotactic rules [*mendi* (hill)]. Marked Basque words were always formed by at least one illegal bigram when measured according to the Spanish vocabulary. Besides, their mean bigram frequency when measured in Spanish fell below the mean log10 frequency of all existing Spanish bigrams as measured from LEXESP (Sebastián-Gallés et al., 2000). In contrast, unmarked Basque words were formed by valid bigram combinations in both languages with mean bigram frequencies falling above the mean log10 Spanish bigram frequency distribution. Spanish words were also split in two sets that were carefully matched between them. One of the Spanish set was assigned as matched control for Basque marked words and the other one was selected as a control for Basque unmarked words. All possible sublexical and lexical factors were equated across and within sets (see **Table 2**).

Procedure

Participants were tested individually in a quiet room using DMDX software (Forster and Forster, 2003) on a 15" monitor set at 90 Hz. Stimuli were presented in lowercase Courier New white letters on a black background. First, a fixation point appeared on the screen for 500 ms followed by the target until participants' response (or for 2500 ms). Feedback was provided only when participants made a mistake. Participants were asked to respond with the right hand to Basque words and with the left hand to Spanish words using a response box. Trial presentation order was randomized across participants. Twenty practice trials were included prior to the experimental trials. The experimental session approximately lasted for approximately 30 min.

RESULTS AND DISCUSSION

Erroneous responses were excluded from the latency analysis as well as responses above or below 2.5 standard deviations from the participants-based and items-based means in each condition (4.80% Balanced Bilinguals, 4.60% Unbalanced Bilinguals, 4.52% Monolinguals). ANOVAs on mean latencies for correct responses and error rates were conducted following a 3 (Group: Balanced Bilinguals, Unbalanced Bilinguals, Monolinguals) × 2 (Language: Spanish, Basque) × 2 (Bigram: Marked, Unmarked) design. Comparisons of the effects were also conducted within and between groups by subtracting the RTs and error rates in Basque trials from the RTs and error rates in the Spanish trials. Mean latencies and error rates are presented in **Table 3** and the effects are plotted in **Figure 1**.

REACTION TIMES

The main effect of Language was not significant [$F1/F2 < 0.85$, $ps > 0.35$]. The main effect of Bigrams was significant [$F1_{(1, 57)} = 171.34$, $p < 0.001$; $F2_{(1, 676)} = 112.23$, $p < 0.001$], suggesting that marked words were recognized faster than

Table 1 | Mean levels of Spanish and Basque language proficiency calculated according to participants' self-ratings (in a 1-to-10 scale).

Language proficiency	Balanced		Unbalanced		Monolinguals	
	Spanish	Basque	Spanish	Basque	Spanish	Basque
Speaking	9.85 (0.46)	9.62 (0.64)	9.88 (0.33)	7.08 (0.89)	9.75 (0.55)	–
Understanding	9.88 (0.33)	9.81 (0.40)	9.92 (0.27)	8.42 (0.86)	9.50 (0.61)	–
Writing	9.69 (0.55)	9.46 (0.81)	9.65 (0.75)	6.92 (1.26)	9.68 (0.47)	–
Reading	9.88 (0.33)	9.81 (0.49)	9.77 (0.65)	8.12 (1.14)	9.48 (0.72)	–
General self-perception	9.81 (0.40)	9.61 (0.50)	9.73 (0.45)	7.54 (0.71)	9.45 (0.76)	–

Standard deviations are provided within parentheses.

Table 2 | Mean values for each sub-lexical, lexical, and semantic factor of the L1 (Spanish) and L2 (Basque) word used split by condition.

	BASQUE		SPANISH	
	Marked	Unmarked	Control marked	Control unmarked
Word frequency	52.00 (114.53)	47.36 (109.53)	44.65 (81.17)	42.56 (74.86)
Word length	6.62 (1.83)	6.81 (2.22)	6.81 (1.81)	6.82 (1.77)
Number of orthographic neighbors	1.42 (1.62)	1.55 (0.35)	1.53 (2.74)	1.69 (3.01)
Age of acquisition	3.22 (0.49)	3.23 (0.50)	3.19 (0.56)	3.19 (0.61)
Word concreteness	4.09 (0.89)	4.12 (0.86)	4.05 (0.81)	4.07 (0.85)
Spanish bigram frequency	1.72 (0.3)	2.97 (0.24)	2.49 (0.30)	2.46 (0.33)
Basque bigram frequency	2.88 (0.18)	2.89 (0.20)		
Number of spanish-implausible bigrams	2.35 (0.93)	0 (0)		

Standard deviations are provided within parentheses.

Table 3 | Mean latencies (in milliseconds) and error rates (in percentage) for words in the four conditions and participant groups for speeded language decision task (Experiment 1).

	Balanced		Unbalanced		Monolingual	
	RT	Error rate	RT	Error rate	RT	Error rate
L2 unmarked	667 (75)	3.35 (2.79)	682 (113)	5.21 (3.82)	689 (128)	8.16 (5.40)
L2 marked	631 (72)	1.97 (1.60)	635 (95)	2.30 (2.24)	564 (87)	1.56 (1.59)
L1 control unmarked	672 (77)	3.01 (1.54)	674 (112)	4.09 (1.90)	603 (102)	3.50 (2.95)
L1 control marked	667 (74)	3.98 (2.14)	667 (113)	4.41 (2.25)	600 (128)	2.88 (2.34)
Unmarked effect	−5 (27)	0.35 (2.46)	8 (30)	1.12 (2.97)	87 (20)	4.66 (3.92)
Marked effect	−36 (16)	−2.00 (1.92)	−32 (31)	−2.12 (2.48)	−36 (10)	−1.32 (0.90)

Standard deviations of the means are provided within parenthesis.

unmarked words. The main effect of group did not reach significance in the analysis by participants, but it was significant in the by-item analysis [$F1_{(2,\ 57)} = 1.72$, $p = 0.19$; $F2_{(2,\ 1352)} = 202.99$, $p < 0.001$]. Critically, the three-way interaction was significant [$F1_{(2,\ 57)} = 36.49$, $p < 0.001$; $F2_{(2,\ 1352)} = 56.13$, $p < 0.001$]. For Basque marked words, all groups tended to respond faster to them than to their corresponding Spanish control words (all ts > 4.5 and ps < 0.001). Furthermore, this markedness effect (i.e., Basque marked words minus Spanish control words) was similar across all groups of participants (all ts < 0.6, ps > 0.55). In contrast, a different pattern emerged for unmarked Basque words. Balanced and Unbalanced Bilinguals responded similarly to unmarked Basque words and to their Spanish controls (all ts < 1.5 and ps > 0.25), while monolinguals took more time to recognize unmarked Basque words than Spanish controls (i.e., an inhibitory effect; [$t_{(19)} = -8.59$, $p < 0.001$]) (see **Table 3**).

ERROR RATES

The statistical analysis on the accuracy data fully replicated the pattern observed in the RTs. The main effect of Language was not significant [$F1/F2 < 0.2$, ps > 0.65] and the main effect of Bigram was significant [$F1_{(1,\ 57)} = 52.23$, $p < 0.001$; $F2_{(1,\ 676)} = 17.51$, $p < 0.001$], showing more errors for unmarked than for marked words. Again, the main Group effect did not reach significance in the by-participants analysis [$F1_{(2,\ 57)} = 1.84$, $p = 0.31$; $F2_{(2,\ 1352)} = 6.64$, $p < 0.005$]. Critically, the three-way interaction was significant [$F1_{(2,57)} = 5.65$, $p < 0.01$; $F2_{(2,\ 1352)} = 5.15$, $p < 0.01$], showing the same pattern of results observed in the RT analysis. In general, participants made fewer errors with

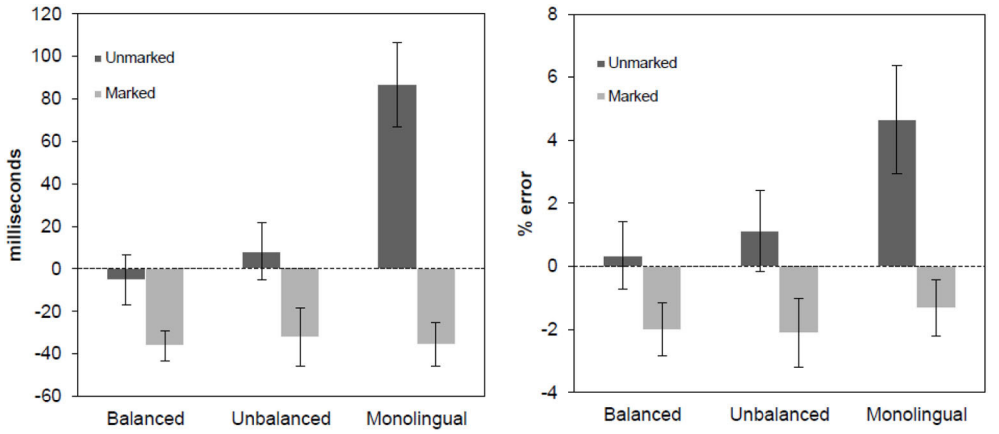


FIGURE 1 | Language effect in reaction times (left panel) and error rates (right panel) for speeded language decision task, separated by marked and unmarked conditions. The effect was obtained subtracting the responses to the Spanish word from the responses to the Basque words. Error bars represent 95% confidence intervals.

Basque-marked words than with Spanish control words (all $ts > 2.5$, $p < 0.05$) and the magnitude of the effects (i.e., the differences between Basque marked words and theirs Spanish control words) did not differ between groups (all $t < 0.15$, $p > 0.25$). For unmarked Basque words, the accuracy rates were similar to that for Spanish control words in the groups of Balanced and Unbalanced Bilinguals ($ts < 2$ and $ps > 0.1$). In contrast, Monolinguals made more errors on unmarked Basque words than on their Spanish controls [$t_{(19)} = -5.32, p < 0.001$].

In general, L2 words that violated the orthotactic rules of Spanish vocabulary (i.e., marked Basque words) were easier to recognize than Spanish words for all groups (faster RTs and lower error rates), suggesting that readers base their decisions regarding the language membership of the words based on orthographic cues. Interestingly, this was true for the two groups of bilinguals, regardless of their clear-cut differences in Basque proficiency, and more strikingly, this was also true for the group of monolinguals with no prior experience with Basque. This result raises a critical question regarding the etiology of this effect. The fact that all participants showed identical markedness effects for L2-specific Basque words suggests that the cognitive processes underlying language discrimination of orthographically-marked words are guided by basic sub-lexical processes associated with the detection of non-native bigram combinations (i.e., the detection of L1-invalid cues; see Vaid and Frenck-Mestre, 2002).

Another critical finding from Experiment 1 helps us qualifying the real cognitive mechanisms leading to efficient language discrimination in bilinguals and monolinguals. Basque words following the Spanish orthotactic rules (i.e., unmarked Basque words) were notably difficult to recognize for monolinguals, but not for bilinguals. This effect of unmarked Basque words that was only present for monolinguals, together with the results observed for marked Basque words across the three groups of participants, suggest that there are two clearly different mechanisms driving language detection depending on the specific orthographic characteristics of the words. First, some form of lexical access seems to determine language detection mechanisms for unmarked words, given the obvious differences in the performance of bilinguals and monolinguals with these stimuli (namely, an inhibitory effect only present in the group of monolinguals, who lack a lexical representation for those items). Second, decisions to L2-marked words seem to be governed by a series of visuo-orthographic processes, rather than by lexical access, given the highly similar performance of all groups with marked Basque words.

In order to better characterize the importance of orthographic cues in bilingual lexical access, and to explore in depth the extent to which visuo-orthographic and lexico-semantic mechanisms determine bilingual visual word recognition we run a second experiment. In Experiment 2 we asked the same groups of participants to perceptually recognize the same Spanish and Basque (marked and unmarked) words in a progressive demasking task. Since correctly completing this task requires retaining the whole strings of letters in memory, a lexically-mediated recognition strategy would yield higher efficiency (shorter reaction times and lower error rates) during the task. Letter strings that have an actual lexical node would be encoded in episodic memory for posterior retrieval more efficiently than letter strings that are not represented in the lexicon. Therefore, we expected that bilinguals would benefit from the presence of such an entry in the lexicon compared to monolinguals. Furthermore, we expected lower activation thresholds of L2 lexical items for balanced bilinguals compared to unbalanced bilinguals reflected in shorter reaction times. Considering the characteristics of the task used in Experiment 2, we predicted that Basque words should take longer to recognize (and lead to higher error rates) than Spanish words for monolinguals and unbalanced bilinguals, but not for balanced bilinguals (who share two L1s). Given the importance of orthographic cues for all groups of participants (as seen in Experiment 1), we predicted that for balanced and unbalanced bilinguals marked Basque words should be recognized faster than unmarked words. In contrast, monolinguals should display now either similar or more difficulty in recognizing words containing letter combinations that are not present in their language, given that the encoding in working memory of letter sequences that

have not been faced beforehand would require a costly perceptual and orthographic analysis of each stimulus.

EXPERIMENT 2: PROGRESSIVE DEMASKING TASK (PDM)

PARTICIPANTS AND STIMULI

These were the same as in Experiment 1.

PROCEDURE

Participants were asked to identify the displayed words as fast and as accurately as possible typing on the keyboard the word they think they read. The experiment was run using the PDM software (Dufau et al., 2008). Trials were composed of target–mask pairs that were consecutively repeated several times. In each trial, the total display time of the stimulus was held constant at 210 ms, and the ratio of the target and mask display durations progressively increased in cycles. In the first cycle, the mask display duration was much longer than the target one (195 and 15 ms, respectively). In the following cycles, the mask display duration decreased and the target display duration increased in a constant way. Participants had to press the spacebar when they had recognized the word, and then type it. Reaction times (RTs) were measured from the initial display of the mask in the first cycle to the button press.

DATA ANALYSIS

Erroneous responses and responses above and below 2.5 standard deviations from the mean of each subject within each condition and of each item were excluded from the analysis of reaction times (2.93% for Balanced Bilinguals, 3.48% for Unbalanced Bilinguals and 3.98% for Monolinguals). The same design from Experiment 1 was followed for the ANOVAs. Mean latencies and error rates are presented in **Table 4** and effects are plotted in **Figure 2** (upper panel).

RESULTS AND DISCUSSION

REACTION TIMES

The main effect of Language was significant [$F1_{(1,\ 57)} = 481.24$, $p < 0.001$; $F2_{(1,\ 676)} = 290.20$, $p < 0.001$], showing that Spanish words were recognized faster than Basque words. The main effect of Bigram was also found in the analysis by participants [$F1_{(1,57)} = 55.59$, $p < 0.001$; $F2_{(1,676)} = 1.69$, $p = 0.21$]. The main effect of Group was significant [$F1_{(2,57)} = 18.59$, $p < 0.001$; $F2_{(2,1352)} = 1982.76$, $p < 0.001$], suggesting that monolinguals in general were slower recognizing words than bilinguals. The three-way interaction was significant [$F1_{(2,57)} = 14.18$, $p < 0.001$; $F2_{(2,1352)} = 5.55$, $p < 0.05$]. All groups responded significantly slower to unmarked Basque words than to the Spanish control words (all ts > 7.5, ps < 0.001), while a different pattern emerged for marked Basque words. While Unbalanced bilinguals and Monolinguals were significantly slower in responding to marked Basque words than to the corresponding Spanish control words (all ts > 6.4 and ps < 0.001), no such difference was found for Balanced bilinguals, who recognized marked Basque words as fast as the Spanish control words [$t_{(19)} = -1.04$, $p > 0.3$]

The differences in the magnitude of the effects between unmarked Basque words and their Spanish control words were also different across all three groups (all ts > 5.5, ps < 0.001), increasing as an inverse function of their proficiency in the language (see **Figure 2**). In contrast, the markedness effect also increased as an inverse function of the participants' proficiency in Basque (all ts > 5.35 and ps < 0.001). Interestingly, the analysis of the magnitude of the effects between marked and unmarked words revealed a facilitative effect for the two bilingual groups [Balanced bilinguals: $t_{(19)} = 4.96$, $p < 0.001$; Unbalanced bilingual: $t_{(19)} = 3.21$, $p < 0.005$], showing that L2-marked words were recognized faster than L2-unmarked words, and an inhibitory effect for the monolingual group [$t_{(19)} = -2.44$, $p < 0.05$], showing that L2-marked words were more difficult to recognize than L2-unmarked words.

ERROR RATES

The main effect of Language was significant [$F1_{(1,\ 57)} = 78.16$, $p < 0.001$; $F2_{(1,\ 676)} = 199.53$, $p < 0.001$], showing that in general Spanish words were recognized more accurately than Basque words. The main effect of Bigram was also significant [$F1_{(1,\ 57)} = 22.58$, $p < 0.001$; $F2_{(1,\ 676)} = 3.01$, $p < 0.05$], as well as the main effect of Group [$F1_{(2,\ 57)} = 33.16$, $p < 0.001$; $F2_{(2,\ 1352)} = 229.05$, $p < 0.001$], showing that monolinguals made more errors than both bilingual groups. Critically, the three-way interaction was significant [$F1_{(2,\ 57)} = 21.20$, $p < 0.001$; $F2_{(2,\ 1352)} = 9.46$, $p < 0.001$], showing a different pattern of the effects for the three types of participants. Not surprisingly, Balanced bilinguals did not show any reliable difference across all conditions (all ts < 1.75 and ps > 0.1), given their high and comparable degree of proficiency in the two languages. In contrast, Unbalanced bilinguals and Monolinguals responded more accurately to Spanish words than to Basque marked and unmarked words (all ts > 2.85 and ps < 0.01), and the error rates decreased as a function of increased proficiency in Basque (all ts > 2.37 and ps < 0.05). Interestingly, the magnitude of the effects between marked and unmarked words revealed an inhibitory effect for the monolingual group [$t_{(19)} = -5.06$, $p < 0.001$], showing that Spanish monolingual participants made more errors typing L2-marked words than L2-unmarked words. No statistical differences were found for any of the bilingual groups [all ts < 1.5, ps > 0.15].

The results of Experiment 2 were clear-cut. Balanced bilinguals took the same amount of time to identify Spanish words and marked Basque words, while in all the other groups and conditions, a generalized identification cost was evident for Basque (L2) words. Also, participants made more errors when typing L2 words than L1 words, but again balanced bilinguals did not show any difference in their accuracy of response to Spanish and Basque words. Interestingly, different markedness patterns emerged for monolinguals as compared to both balanced and unbalanced bilinguals. Monolinguals took more time and made more errors in recognizing L2-marked than unmarked words, but the opposite pattern was found for both types of bilinguals who showed faster responses for L2-marked than unmarked words and no significant differences in terms of accuracy. Thus, these results suggest that different mechanisms or strategies are involved in the recognition of strings of letters that include legal and illegal bigram combinations, depending on the existence of lexical representations associated with the target strings (i.e., a lexical strategy vs. an orthographic-to-phonological working memory strategy).

Table 4 | Mean latencies (in milliseconds) and error rates (in percentage) for words in the four conditions and participant groups for progressive demasking task (Experiment 2).

	Balanced		Unbalanced		Monolingual	
	RT	Error rate	RT	Error rate	RT	Error rate
L2 unmarked	1404 (234)	2.65 (2.19)	1438 (225)	4.71 (2.62)	2064 (305)	16.88 (10.42)
L2 marked	1351 (227)	2.68 (2.44)	1481 (228)	4.21 (2.79)	2067 (306)	22.88 (11.77)
L1 control unmarked	1352 (223)	2.03 (1.80)	1343 (214)	1.91 (1.69)	1487 (236)	2.65 (1.96)
L1 control marked	1343 (216)	2.74 (1.90)	1330 (191)	2.47 (1.44)	1458 (224)	2.71 (1.61)
Unmarked effect	52 (30)	0.62 (1.60)	138 (26)	2.79 (2.55)	557 (137)	14.24 (3.03)
Marked effect	8 (35)	−0.05 (2.04)	108 (75)	1.74 (2.69)	609 (140)	20.18 (10.89)

Standard deviations of the means are provided within parenthesis.

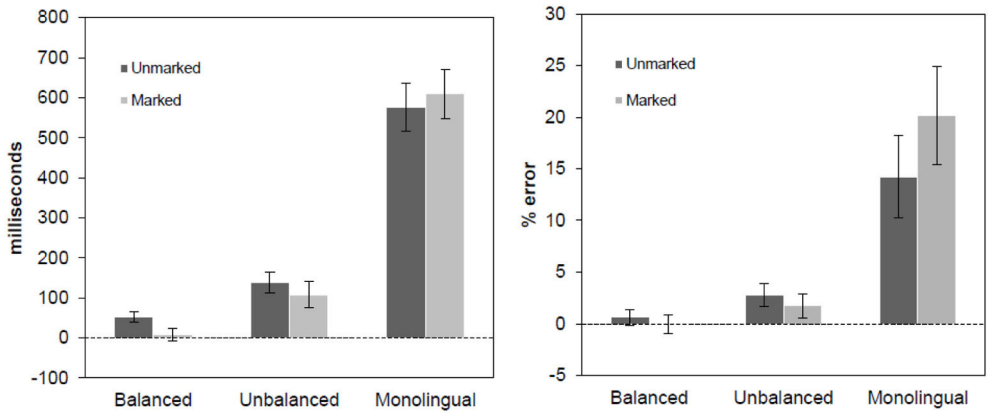


FIGURE 2 | Language effect in reaction times (left panel) and error rates (right panel) for PDM task, separated by markedness conditions. The effect was obtained subtracting the responses to the Spanish word from the responses to the Basque words. Error bars represent 95% confidence intervals.

Together, these results suggest that (1) bilingual participants followed a lexical-search strategy when recognizing marked and unmarked words, while monolinguals followed a different encoding strategy, and that (2) L2-marked words help bilingual lexical access, leading to advantageous word identification as compared to words that orthographically speaking can also belong to their L1. This suggests that bilinguals rely on both sub-lexical and lexical information during multilingual perceptual identification of words.

GENERAL DISCUSSION

The main goal of the present study was to investigate how the sub-lexical characteristics of the words from bilinguals' two languages influence different stages of the visual word recognition process. Three groups of participants (balanced bilinguals, unbalanced bilinguals, monolinguals) were tested in a language decision task and in a progressive demasking task. Materials consisted of a selection of Spanish (L1) and Basque (L2) words. Crucially, L2 words could follow L1 orthotactic rules (i.e., language-unspecific orthography; L2-unmarked words) or violate L1 orthotactic rules in terms of the corresponding bigram frequencies (i.e., language-specific orthography; L2-marked words). Results showed that L2-marked and unmarked words were recognized differently depending on the task demands and on the participants' linguistic profile. When the task required explicitly focus on the language tag (Experiment 1), all group of participants showed strong markedness effects (namely, an advantage in the recognition of L2-marked words) independently of their L2 knowledge and proficiency. Language-specific orthography speeded up participants' language decisions. However, when the task required participants to fully identify the strings (Experiment 2), the L2-markedness advantage only emerge for bilingual participants. Additionally, these effects were clearly modulated as a function of bilinguals' L2 lexical knowledge and proficiency.

As seen in Experiment 1, all participants seem to have based their language decisions on the existing sub-lexical cues, as reflected by the generalized benefit for L2-marked words (which were recognized even faster than L1 words). Critically, this effect was present for all types of participants, regardless of their knowledge of the L2 and their proficiency in that language. At first glance, this result could be taken as evidence supporting the sub-lexical strategy that has been suggested to guide bilinguals' language identification (see Vaid and Frenck-Mestre, 2002). Furthermore, results from Experiment 1 could be taken as a confirmation of the existence of tight links between sub-lexical information and language membership (see Van Kesteren et al., 2012). However, a closer look at the effects found in

Experiment 1 for L2-unmarked words suggests that the sub-lexical strategy is not the only mechanism at play during language discrimination. When such orthotactic cues were not available to participants (namely, L2-unmarked words), all participants (bilinguals and monolinguals) seem to have followed a lexical-search strategy, given that they did not have any cue other than the match between the printed string and their known lexical forms to assign the language. Bilinguals performed notably well with L2-unmarked words, given the existence of L2 lexical representations associated with these strings, while this was not the case for monolinguals. Monolinguals took more time to recognize L2-unmarked words, most probably due to an intensive and fruitless lexical search for those items.

These two different strategies (lexical vs. sub-lexical) are correctly accommodated by current models of bilingualism (i.e., BIA+, Dijkstra and Van Heuven, 2002; and the extension of the BIA+, Van Kesteren et al., 2012), insofar they suggest that language membership can be accessed (1) once the lexical representations are activated, and also (2) directly from sub-lexical levels of processing. Our results help to better defining the specific situations in which these two routes are followed, clarifying the specific scenarios in which the sub-lexical strategy may be useful. On one hand, all readers seem to follow the sub-lexical strategy for L2-marked words. Hence, bilinguals seem to base their judgments for L2-marked words on a sub-lexical strategy that is also shared by monolinguals. On the other hand, when no orthographic cues are available (L2-unmarked words) participants follow a lexical strategy based on the identification of the correspondent word form in the lexicon (and hence the differences between monolinguals and bilinguals). However, according to this lexical search strategy, a clear modulation of the effects for L2-unmarked words would have also been expected within the two bilingual samples, given their obvious L2-proficiency differences. Nonetheless, this proficiency effect for L2-unmarked words was absent in Experiment 1, since the effect was highly similar for both groups of bilinguals. We tentatively proposed that the language decision task might not be sensitive enough to capture these subtle differences based on participants' L2 proficiency, and Experiment 2 confirmed this intuition.

Experiment 2 qualified, complemented and extended the observations from Experiment 1. First, the results from the progressive demasking task suggest that when language membership assignment is not the main aim of the task, participants mainly rely on their lexicon, partially abandoning the sub-lexical strategy followed in Experiment 1 and focusing on their lexical knowledge. Balanced bilinguals performed similarly with L1 and L2 words, while unbalanced bilinguals were significantly slower and made more errors for L2 words than for L1 words. Besides, monolinguals were markedly slow and inaccurate in identifying Basque (unknown) words. Hence, in contrast to Experiment 1, Experiment 2 clearly showed a graded pattern of effects associated with proficiency.

Critically, all bilinguals (balanced and unbalanced) showed a benefit in their speed of recognition for L2-marked words as compared to L2-unmarked words, suggesting that even in a task in which language membership assignment is not required, early detection of the language through a sub-lexical analysis of the words aids lexical access. On the basis of their statistical regularities, L2-unmarked words would initially activate Spanish and Basque lexical candidates, while L2-marked words would provide bilingual readers with a critical cue exclusively pointing to the Basque vocabulary (see Grainger and Beauvillain, 1987; Schwartz et al., 2007). Obviously, these cues would not be helpful for monolinguals, given the absence of a Basque lexicon. These results fit well with the postulates of the extension of the BIA+ model proposed by Van Kesteren et al. (2012), who suggested that information regarding language-specific sub-lexical information aid language detection. Importantly the present results extend their claims by showing that even in a context in which assignment of language membership is not required, sub-lexical cues aid lexical search by inhibiting lexical representations from the non-target language or aiding the selection of the target language, thus facilitating lexical access. In the absence of these orthographic cues, the multiplicity of activated lexical candidates from the L1 and the L2 results in a high degree of dispersion of the activation, leading to an enhanced difficulty in selecting the correct representation from the lexicon. These results fit well with some of the mechanisms proposed in the BIA+ model (Dijkstra and Van Heuven, 2002), demonstrating that when a printed word is presented to a bilingual, both languages would be initially activated (non-selective access), but as proposed by Van Kesteren et al., in the presence of orthographic cues the sub-lexical features would allow for certain degree of selective lexical access. Moreover, our results also fit well with the dual-route account specified in the BIA+ extended model. According to Van Kesteren et al., language membership information could be accessed through the retrieval of the lexical information of the words, or directly via sub-lexical information of the letter strings. When the goal of the task is to detect language membership (e.g., language decision tasks; Experiment 1), task-related decisions could be made based on the direct links established between sub-lexical nodes and language membership, making full lexical access unnecessary. That is, L2-marked words could be detected just following a sub-lexical strategy. However, when lexical access is required to correctly perform the task (e.g., word identification tasks; Experiment 2), the sub-lexical route remains effective, but decisions are also mediated by a lexical search strategy. That is, L2-marked words in a word identification task would simultaneously activate both the lexical and sub-lexical routes, which are interconnected following interactive activation principles, thus facilitating bilingual single word recognition.

It is well known that sub-lexical orthographic regularities of the words have a direct impact in the way in which monolingual readers decipher the written code, as shown by multiple studies demonstrating the impact of bigram frequencies in visual word recognition (e.g., Whitney, 2001; Grainger and Van Heuven, 2003; Whitney and Cornelissen, 2008; Dandurand et al., 2011). However, to date little is known about the impact of these orthographic regularities in bilingual reading, and moreover, about the manner in which these regularities can be unconsciously used as

access cues to the bilingual lexicon. The results reported in this article demonstrate the importance of sub-lexical orthographic features in bilingual reading by showing a high degree of sensitivity of bilingual readers to language-specific bigram combinations that is strikingly different from the pattern seen in monolingual readers under the appropriated experimental contexts. In summary, we have shown that L2-marked words are always faster to recognize than L2-unmarked words for individuals who are immersed in bilingual contexts (but not for monolinguals), independently of the task demands. Besides, we have shown that the reliance on sub-lexical information seems to depend on the specific nature of the task and, more importantly, on the proficiency of the participants in the second language, in spite of their permanent exposure to the two languages in a naturalistic context. The current results demonstrate the existence of (at least) two possibly interconnected strategies during bilingual lexical access: a sub-lexical visuo-orthographic stage that is highly sensitive to the specific language cues, and a lexical search strategy. Thus, the differences between the orthotactic rules of two languages that share the same script are extremely important for language detection, and ultimately for lexical access in bilingual contexts.

ACKNOWLEDGMENTS

This research was partially supported by grants CSD2008-00048, PSI2012-31448 and PSI2012-32123 from the Spanish Government, and ERC-AdG-295362 and FP7-SSH-2013-1-GA613465 from the European Research Council.

REFERENCES

Carreiras, M., Perea, M., and Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: cross-task comparisons. *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 857–871. doi: 10.1037/0278-7393.23.4.857

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. doi: 10.1037/0278-7393.32.5.1057

Dandurand, F., Grainger, J., Duñabeitia, J. A., and Granier, J. P. (2011). On coding non-contiguous letter combinations. *Front. Psychol.* 2:136. doi: 10.3389/fpsyg.2011.00136

Davis, C. J., and Perea, M. (2005). BuscaPalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behav. Res. Meth.* 37, 665–671. doi: 10.3758/BF03192738

Dijkstra, T., and Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Biling. Lang. Cogn.* 5, 175–197. doi: 10.1017/S1366728902003012

Dimitropoulou, M., Duñabeitia, J. A., and Carreiras, M. (2011). Two words, one meaning: evidence of automatic co-activation of translation equivalents. *Front. Psychol.* 2:188. doi: 10.3389/fpsyg.2011.00188

Dufau, S., Stevens, M., and Grainger, J. (2008). Windows Executable Software for the Progressive Demasking Task (PDM). *Behav. Res. Methods* 40, 33–37. doi: 10.3758/BRM.40.1.33

Duñabeitia, J. A., Avilés, A., and Carreiras, M. (2008). NoA's Ark: influence of the number of associates in visual word recognition. *Psychon. Bull. Rev.* 15, 1072–1077. doi: 10.3758/PBR.15.6.1072

Duñabeitia, J. A., Dimitropoulou, M., Uribe-Etxebarria, O., Laka, I., and Carreiras, M. (2010a). Electrophysiological correlates of the masked translation priming effect with highly proficient simultaneous bilinguals. *Brain Res.* 1359, 142–154. doi: 10.1016/j.brainres.2010.08.066

Duñabeitia, J. A., Perea, M., and Carreiras, M. (2010b). Masked translation priming effects with highly proficient simultaneous bilinguals. *Exp. Psychol.* 57, 98–107. doi: 10.1027/1618-3169/a000013

Forster, K. I., and Forster, J. C. (2003). DMDX: a windows display program with millisecond accuracy. *Behav. Res. Meth. Ins. C.* 35, 116–124. doi: 10.3758/BF03195503

Francis, W. S., Tokowicz, N., and Kroll, J. F. (2014). The consequences of language proficiency and difficulty of lexical access for translation performance and priming. *Mem. Cogn.* 42, 27–40. doi: 10.3758/s13421-013-0338-1

Grainger, J., and Beauvillain, C. (1987). Language blocking and lexical access in bilinguals. *Q. J. Exp. Psychol.* 39A, 295–319. doi: 10.1080/14640748708401788

Grainger, J., and Segui, J. (1990). Neighborhood frequency effects in visual word recognition: a comparison of lexical decision and masked identification latencies. *Percept. Psychophys.* 47, 191–198. doi: 10.3758/BF03205983

Grainger, J., and Van Heuven, W. (2003). "Modeling letter position coding in printed word perception," in *The Mental Lexicon*, ed P. Bonin (New York, NY: Nova Science Publishers), 1–24.

Jacobs, A. M., and Grainger, J. (1994). Models of visual word recognition—sampling the state of the art. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 1311–1334. doi: 10.1037/0096-1523.20.6.1311

Jacobs, A. M., Rey, A., Ziegler, J. C., and Grainger, J. (1998). "MROM-p: an interactive activation, multiple readout model of orthographic and phonological processes in visual word recognition," in *Localist Connectionist Approaches to Human Cognition*, eds J. Grainger and A. M. Jacobs (Mahwah, NJ: Erlbaum), 147–188.

Kuipers, J. R., and Thierry, G. (2010). Event-related brain potentials reveal the time-course of language change detection in early bilinguals. *Neuroimage* 50, 1633–1638. doi: 10.1016/j.neuroimage.2010.01.076

Lemhöfer, K., Dijkstra, T., Schriefers, H., Harald, R., Grainger, J., and Zwitserlood, P. (2008). Native language influences on word recognition in a second language: a megastudy. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 12–31. doi: 10.1037/0278-7393.34.1.12

Lemhöfer, K., Koester, D., and Schreuder, R. (2011). When bicycle pump is harder to read than bicycle bell: effects of parsing cues in first and second language compound reading. *Psychon. Bullet. Rev.* 18, 364–370. doi: 10.3758/s13423-010-0044-y

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Appl. Linguist.* 24, 578–596. doi: 10.1093/applin/amm056

Perea, M., Duñabeitia, J. A., and Carreiras, M. (2008). Masked associative/semantic and identity priming effects across languages with highly proficient bilinguals. *J. Mem. Lang.* 58, 916–930. doi: 10.1016/j.jml.2008.01.003

Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., and Carreiras, M. (2006). E-Hitz: a word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behav. Res. Meth.* 38, 610–615. doi: 10.3758/BF03193893

Pliatsikas, C., and Marinis, T. (2013). Processing of regular and irregular past tense morphology in highly proficient second language learners of English: a self-paced reading study. *Appl. Psycholinguist.* 34, 943–970. doi: 10.1017/S0142716412000082

Schwartz, A. I., Kroll, J. F., and Diaz, M. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Lang. Cogn. Process.* 22, 106–129. doi: 10.1080/01690960500463920

Sebastián-Gallés, N., Martí, A., Carreiras, M., and Cuetos, F. (2000). *LEXESP: Una Base de Datos Informatizada del Español.* Barcelona: Universitat de Barcelona.

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. doi: 10.1073/pnas.0609927104

Thomas, M. S. C., and Allport, A. (2000). Language switching costs in bilingual visual word recognition. *J. Mem. Lang.* 43, 44–66. doi: 10.1006/jmla.1999.2700

Vaid, J., and Frenck-Mestre, C. (2002). Do orthogonal cues aid language recognition? A laterality study with French–English bilinguals. *Brain Lang.* 82, 47–53. doi: 10.1016/S0093-934X(02)00008-1

Van Kesteren, R., Dijkstra, T., and de Smedt, K. (2012). Markedness effects in Norwegian–English bilinguals: task-dependent use of language-specific letters and bigrams. *Q. J. Exp. Psychol.* 65, 2129–2154 doi: 10.1080/17470218.2012.679946

Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G., Shiffrin, R. M., Van Rijn, H., and Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cogn. Psychol.* 48, 332–367. doi: 10.1016/j.cogpsych.2003.08.001

Westbury, C., and Buchanan, L. (2002). The probability of the least likely non-length-controlled bigram affects lexical decision reaction times. *Brain Lang.* 81, 66–78. doi: 10.1006/brln.2001.2507

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: the SERIOL model and selective literature review. *Psychon. Bull. Rev*. 8, 221–243. doi: 10.3758/BF03196158

Whitney, C., and Cornelissen, P. (2008). SERIOL reading. *Lang. Cogn. Process.* 23, 143–164. doi: 10.1080/01690960701579771

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 February 2014; accepted: 22 April 2014; published online: 13 May 2014.
Citation: Casaponsa A, Carreiras M and Duñabeitia JA (2014) Discriminating languages in bilingual contexts: the impact of orthographic markedness. Front. Psychol. ***5****:424. doi: 10.3389/fpsyg.2014.00424*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 27 October 2014
doi: 10.3389/fpsyg.2014.01203

Native-likeness in second language lexical categorization reflects individual language history and linguistic community norms

Benjamin D. Zinszer [1,2] ****, Barbara C. Malt*** [3] ***, Eef Ameel*** [4] ***and Ping Li*** [1]

[1] *Department of Psychology, Center for Language Science, Pennsylvania State University, University Park, PA, USA*
[2] *Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA*
[3] *Department of Psychology, Lehigh University, Bethlehem, PA, USA*
[4] *Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium*

Edited by:
Vicky Chondrogianni, University of Edinburgh, UK

Reviewed by:
Anthony Shook, Northwestern University, USA
Robert Nelson, University of Alabama, USA

****Correspondence:***
Benjamin D. Zinszer, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA
e-mail: bzinszer@gmail.com

Second language learners face a dual challenge in vocabulary learning: First, they must learn new names for the 100s of common objects that they encounter every day. Second, after some time, they discover that these names do not generalize according to the same rules used in their first language. Lexical categories frequently differ between languages (Malt et al., 1999), and successful language learning requires that bilinguals learn not just new words but new patterns for labeling objects. In the present study, Chinese learners of English with varying language histories and resident in two different language settings (Beijing, China and State College, PA, USA) named 67 photographs of common serving dishes (e.g., cups, plates, and bowls) in both Chinese and English. Participants' response patterns were quantified in terms of similarity to the responses of functionally monolingual native speakers of Chinese and English and showed the cross-language convergence previously observed in simultaneous bilinguals (Ameel et al., 2005). For English, bilinguals' names for each individual stimulus were also compared to the dominant name generated by the native speakers for the object. Using two statistical models, we disentangle the effects of several highly interactive variables from bilinguals' language histories and the naming norms of the native speaker community to predict inter-personal and inter-item variation in L2 (English) native-likeness. We find only a modest age of earliest exposure effect on L2 category native-likeness, but importantly, we find that classroom instruction in L2 negatively impacts L2 category native-likeness, even after significant immersion experience. We also identify a significant role of both L1 and L2 norms in bilinguals' L2 picture naming responses.

Keywords: lexical categorization, lexical semantics, bilingualism, immersion, language learning

INTRODUCTION

Second language acquisition research has often highlighted the role of learners' language history as a strong predictor of ultimate second language (L2) attainment in syntax and phonology (e.g., Flege, 1987; Johnson and Newport, 1989). Variables of interest have typically included age of acquisition (AOA) and length of residence (LOR) in a second language environment, which have good predictive value for proficiency in syntax and phonology. The roles of these predictors in lexical acquisition, however, have not been as clear when measured through the lens of vocabulary size (e.g., Snow and Hoefnagel-Höhle, 1978) or brain responses to word stimuli (see Weber-Fox and Neville, 1996; Ojima et al., 2011; Granena and Long, 2012 for several perspectives).

A closer examination of lexical semantics reveals, though, that the development of the lexicon may be more analogous to that of syntax and phonology than such divergent outcomes suggest. Recent research in lexical categorization has moved beyond the size of learners' vocabularies and investigated more subtle aspects of word knowledge such as lexical category boundaries in both native and L2 speakers of a language. The studies reviewed below have found significant variation in lexical categorization patterns among native speakers, simultaneous bilinguals, and sequential bilinguals as a function of predictors such as age of onset, language learning experience, and usage patterns. In this paper we examine determinants of L2 lexical acquisition in more detail, with emphases on L2 immersion experience and its interaction with both individual bilinguals' language histories and the word use patterns of the linguistic communities in which both first and second language are acquired.

LEXICAL CATEGORIZATION

Decades of research have indicated differences in lexical categorization across languages (such as the seminal comparison of color categories by Landar et al., 1960), extending beyond abstract domains to concrete domains such as furniture, clothing, and household storage and serving vessels, and observed across Spanish, English, Chinese, Dutch, French, Russian, and more (Graham and Belnap, 1986; Malt et al., 1999, 2003; Ameel et al., 2005; Pavlenko and Malt, 2011; see Malt and Majid, 2013 for review).

These differences mean that to use words as a native speaker does, language learners must acquire non-obvious, language-specific ways of generalizing names to new objects. For native speakers, fine-tuning of lexical categories may begin in infancy, but it continues beyond childhood, at least up to 14 years of age (Ameel et al., 2008), reflecting the significant challenge in language acquisition that word learning poses, even for monolinguals (see also Bowerman and Levinson, 2001). Developing adult, native-like boundaries between close competitor names requires attention to an increasing number of features of an object over time (Ameel et al., 2008). For example, no single concrete or abstract feature is sufficient to isolate members of the English category *bottle* from the set of 60 common household containers used by Malt et al. (1999). Instead, an interplay between features such as shape (typically cylindrical), material (plastic or glass), and function (containment of a fluid) define this broad category of container-like objects.

Learners of a second language, including children who acquire two languages simultaneously, are thus faced with a major incongruity between languages. For example, Chinese (referring to Mandarin Chinese throughout this paper) and English differ in the principal features by which containers are categorized. Native Chinese speakers use *píngzi* for tall, transparent beverage containers (like a 20 oz soft drink) and *guàn* for shorter, rounder, and more extended in volume, containers (like a 12 oz soft drink), analogous to the English categories *bottle* and *can* respectively. However, the relative priority of material (plastic or metal) and shape (height and roundness) as defining features differs between Chinese categories and English categories. A tall, metallic container for shaving cream may be called *píngzi* in Chinese but *can* in English, violating the ostensive translation relationships for *píngzi-bottle* and *guàn-can.*

Lexical categorization is a valuable tool for identifying variation in lexical semantic mappings among speakers, and with this more sensitive measure of lexical semantic variation, second language lexical proficiency may no longer be sufficiently described by the accumulation of a list of words as tested by most picture naming, lexical decision, and fluency tasks. Instead, lexical semantic mappings are more precisely probed when many similar objects are named, which allows inferences about the boundaries of a given speaker's lexical category. For instance, the researcher can examine which drinking vessels are named *cup* and which similar objects receive a different name (such as *mug* or *glass*) by a speaker.

Recent work has investigated whether and how bilinguals can maintain native-like lexical semantic representations in each language despite these differences. Ameel et al. (2005) tested simultaneous Dutch–French bilinguals on the names of common containers and serving dishes. Significant influences of both Dutch and French mappings were measured in the bilinguals' categorization patterns for both languages, and the differences between lexical categories in the bilinguals' Dutch and French were significantly smaller than the differences between monolinguals of each language. In effect, the simultaneous bilinguals partially converged across the two languages. They achieved this convergence by shifting category centroids in each language toward one another for greater consistency between approximate translation equivalents and reducing the number of features used to define category boundaries (Ameel et al., 2009). As such, convergence produces more similar lexical categories in each language and minimizes the conflict faced by the simultaneous bilinguals in organizing the objects into named categories. Sequential bilinguals also show similar trends toward convergence (Pavlenko and Malt, 2011; Malt et al., under review). The accumulating findings in lexical categorization behavior of simultaneous and sequential bilinguals are highly suggestive of a dynamic representation for lexical semantics, mutually influenced by both languages, susceptible to change well into adulthood.

These cross-language transfer and convergence effects can be thought of in terms of how exposure to one language might change mappings from objects' representative features to words in the other language of the bilingual speaker. Theoretical models of lexical semantic representation, such as Van Hell and De Groot's (1998) Distributed Feature Model describe a set of underlying features whose combination may be used to define lexical concepts by linking these features to a lexical node. Models that use feature-based representations have been further adapted to accommodate broader asymmetry between languages (Dong et al., 2005) and the relative salience of different features in bilingual categorization (Ameel et al., 2009).

At least two computational models have attempted to simulate bilingual lexical categorization (Zinszer et al., 2011; Fang et al., 2013), drawing on connectionist architecture to translate language-specific mappings into training parameters for lexical nodes and high-dimensional semantic representations. These models are consistent with previous connectionist models of monolingual word learning (such as McClelland and Rogers, 2003) that rely on distributed feature representations to reproduce semantic category hierarchies (e.g., *sunfish* belongs to *fish*, which belongs to *animals*, all of which differ from *plants*) as a result of feature overlap between exemplars.

Although there are only a few quantitative accounts of bilingual lexical categorization, a number of likely predictors for development of lexical categories are apparent from the broader study of second language acquisition. The extent of L2 immersion, age of second language onset, time spent learning the second language in a formal setting (classroom training), and patterns of language use (the extent to which the languages are intermixed in use) all appear to be involved in non-native learners' degree of success in learning a second language. Further, because name choice for an object may vary across speakers (e.g., Malt et al., 1999) the categorization norms of a linguistic community are an important means of quantifying a language learning environment, describing the variety of lexical semantic mappings used by native speakers in that community.

While many studies in second language acquisition explore the influence of language history variables on lexical learning, fewer studies have evaluated a combination of such variables simultaneously and properly controlled for interaction among the variables and statistical obstacles to measuring effects of variables individually, as outlined by Stevens (2006). None of the research to date has simultaneously related all of these variables to lexical categorization as a measure of word learning. We now consider these

variables and how they may impact L2 development of lexical categorization in more detail.

SECOND LANGUAGE IMMERSION

The value of L2 immersion is uncontroversial in second language acquisition research with respect to many components of L2 acquisition. Recent findings in lexical categorization suggest that as in other domains of language acquisition, native-like L2 lexical categorization is supported by L2 immersion. Malt and Sloman (2003) measured the English lexical categorization of 68 bilinguals (including 15 Chinese–English from various dialect backgrounds) immersed in an English environment by asking them to name pictures of common household containers, comparing the name distributions among the bilinguals to those of native English monolinguals. The bilingual participants had varying levels of English proficiency, years of English study, ages of English acquisition, and durations of English immersion. When contrasted against the other language history variables, time spent in the immersion environment was a significant predictor for the acquisition of native-like lexical semantic mappings. Immersion accounted for the greatest proportion of the variance in participants' L2 native-likeness when entered into a multiple regression alongside years studying L2, suggesting its relative importance above formal language training. Further, age of onset and age of immersion effects were completely removed when regressed alongside length of immersion, highlighting the confounding relationships between these variables and the importance of immersion duration as a confound of age effects (Malt and Sloman, 2003).

Within-category variation arises constantly as part of the natural environment, as one may have occasion to sit in several different *chairs* each day and drink from a variety of *cups*. However, classroom learning includes little exposure to the within-category variation necessary to acquire native-like lexical semantics. Consequently, immersed learners are likely to follow different developmental trajectories than non-immersed learners, as their respective language inputs differ fundamentally in the lexical semantic domain. Additionally, aspects of language history interact or confound with immersion experience, as described in Malt and Sloman's (2003) study above. Understanding other learning variables in concert with immersion may offer a novel perspective on L2 lexical semantic development pre- and post-immersion.

AGE OF L2 ONSET

Age of second language onset as a predictor of eventual second language attainment remains a controversial topic, as evidence for and against a sensitive period for language acquisition is weighed alongside varying levels of other confounding age-related variables (such as years of L2 exposure, motivation, and socialization; see a recent review in Li, 2014). Age effects measured in lexical development by vocabulary size (e.g., the Peabody Picture Vocabulary Test) and translation tasks suggest that older learners may be at an advantage relative to early childhood learners (Snow and Hoefnagel-Höhle, 1978). This effect may arise in part because adults already have existing lexical semantic representations on which to base L2 word learning. One recent ERP study supports this later-is-better advantage for native-likeness of semantic processing (Ojima et al., 2011) while another ERP study (Granena and Long, 2012) indicates an advantage for earlier ages of onset lexical acquisition.

However, these tests do not account for between-language variation in lexical semantic mappings and may overlook non-native word uses by older speakers who rely on direct translation for L2 learning. The relationship between age effects and native-like lexical categorization performance is not entirely clear. Although Malt and Sloman (2003) found a weakly negative effect for later ages of L2 onset, this effect vanished after controlling for immersion. Further, other recent findings have suggested that earlier introduction of L2 may lead to reduced native-likeness of lexical semantic mappings in both L1 and L2. Very early onset Russian–English bilinguals show relatively less similarity to either L1 or L2 norms when speaking L1 compared to their later-onset peers who showed more stable influence of each language over their L1 production (Pavlenko and Malt, 2011). The very early onset bilinguals' unique category patterns may arise from incomplete acquisition of L1 or interference of L2 in the acquisition of L1 patterns. One possible explanation is that L1–L2 interaction dramatically increases in earlier ages of onset, supported by recent computational models (Zinszer and Li, 2010; Li and Zhao, 2013; see also articles in a special issue on computational modeling, ed. Li, 2013) which have demonstrated that prior entrenchment of L1 representations may produce age effects which resemble a sensitive period and that lexical semantic representations are more integrated between languages for early onset learning, while the languages are organized relatively independently for later-onset learners (Li and Zhao, 2013).

The possible departure from conventional "earlier is better" wisdom about age of onset raises questions about whether simultaneous bilinguals are unique in their degree of convergence between languages. If late bilinguals show diminished convergence, more native-like representations may be learnable in both L1 and L2 independently, even when marginal cross-language transfer is observable.

L2 CLASSROOM INSTRUCTION

Malt and Sloman's (2003) study of L2 English learners found that formal training in English prior to immersion offered no predictive power after accounting for years of L2 immersion. Based on this result, L2 training would seem to have minimal value for acquiring native-like L2 lexical semantic mappings. However, some degree of successful L2 lexical semantic remapping has been observed in non-immersed learners with sufficiently advanced L2 education. Chinese students in their third year of undergraduate study as English majors demonstrated significantly higher L2 native-likeness in semantic similarity judgments than a first-year cohort (Dong et al., 2005).

The latter result does not strongly contradict the Malt and Sloman (2003) finding, however, in that the Chinese students of English at both levels still relied primarily on their native Chinese semantics when making English judgments, showing greater similarity to the monolingual Chinese speakers than to English–Chinese bilinguals (native English speakers). Both the

improvement toward slightly more native-like English associations and the general bias toward Chinese semantics are reflected in the learners' significant convergence, producing semantic similarity judgments that were more similar across languages than the judgments between the Chinese monolinguals and English–Chinese bilinguals. For these sequential bilingual learners, language systems interacted to allow a small degree of transfer of learned L2 mappings onto L1 while never overcoming the overall L1-likeness of the representations in both languages. Thus the role of classroom experience in acquiring native-like L2 lexical categorization deserves more scrutiny.

LANGUAGE USE CONTEXT

The type of language experience gained in an immersion environment can vary substantially among bilinguals. Simultaneous bilinguals, such as those in Ameel et al.'s (2005) study are often immersed in an environment that involves frequent input from speakers of both languages. Sequential bilinguals may transition from a monolingual L1 environment to a new language environment where most speakers are monolinguals of L2. In this new environment, L1 use may be limited to a social or familial community, and L2 use may be primarily for work or business.

The monolingual or bilingual context of the language environment or the extent to which speakers switch between languages changes the degree of cognitive control necessary for language production. Specifically, highly bilingual environments raise the potential for frequent code-switching and increase activation of the non-target language, which must then be actively inhibited from production (Green and Abutalebi, 2013). This effect arises from the persistent simultaneous activation of languages (see Kroll et al., 2006, 2012 for reviews) and creates the possibility that each language may be susceptible to change through retrieval induced reconsolidation (Wolff and Ventura, 2009). In retrieval induced reconsolidation (see Forcato et al., 2007), all active representations are adjusted during access by the current input, even if not selected. Because lexical semantic mappings draw on shared cross-language conceptual representation (Van Hell and De Groot, 1998), production of one language may result in reshaping of the other, particularly when both languages are highly active in bilingual environments with more frequent code-switching.

Evidence supporting the view of language change through use can be found in a recent study of phonological accent in the native language. De Leeuw et al. (2010) identified code-switching as a significant predictor in the extent to which the first language phonological system was preserved for bilinguals immersed in an L2 environment. Specifically, greater time spent in L1 environments that inhibited code-switching (such as written correspondence and professional settings) was a significant predictor of L1 stability, while time spent in L1 environments that were permissive of code-switching (e.g., among family and friends) was not associated with preservation of L1 phonology. De Leeuw et al.'s (2010) finding is highly suggestive of the role of language use in regulating the contact and transfer between L1 and L2. Such findings may be relevant to the observation of substantial convergence in simultaneous bilinguals' lexical categorization behavior found by Ameel et al. (2005). Although Ameel et al. (2005) did not directly measure the incidence of intra-sentential code-switching in this setting, the highly bilingual environment is one in which code-switching is more likely to occur.

The contrast observed in Pavlenko and Malt's (2011) early and childhood bilinguals may also reflect the influence of contexts of language use. The early bilinguals in their study (age of L2 onset 6 years or earlier) reportedly participated in a much more fluid bilingual environment from the outset than the child bilinguals (age of L2 onset 8–15 years). These patterns of use are confounded with the age of onset and incomplete L1 acquisition effects and may explain the differences between these groups in native-likeness of L1 and L2. Later-onset of L2 correlates with more discrete separation between language environments and therefore relatively more native-likeness, even as cross-language influence begins to appear.

LINGUISTIC COMMUNITY NORMS

Because native, monolingual speakers of a language also show significant variation in lexical categorization patterns, even monolingual infants acquiring their native language are exposed to variable input for many objects' names. In the relatively familiar domain of household containers, Malt and Sloman (2003) found a broad range of native speaker agreement levels across objects, with the dominant name being produced by as few as 43% of native speakers for some objects (the remaining 57% divided between two or more subordinate names) and 100% agreement for others. In effect, immersed learners are exposed to an array of potential names for many objects, and for some objects the most dominant or native-like name arises in only a minority of encounters (native agreement levels below 50%). L2 learners are thus challenged with determining to *which* of several new categories an object is best suited. This ambiguity results in a many-to-many mapping problem for a single object. For example, a particular serving vessel may be called *diézi* by 70% of Chinese speakers and *pánzi* by 30% of Chinese speakers. Both names may be translated as *dish* or *plate* in English, and *diézi* has the further possible translation of *saucer*. In effect, the Chinese–English bilingual may encounter at least five unique categories of varying fitness for this object from native speakers of the two languages.

As we have discussed earlier, bilinguals' lexical categorization patterns in either language are, indeed, jointly predicted by the native (monolingual) patterns of the two languages (Ameel et al., 2005, 2009; Pavlenko and Malt, 2011; Malt et al., under review). At the earliest stages of learning, before they develop sufficiently elaborated L2 representations, L2 learners draw heavily on L1 representations for production (see the Unified Competition Model of MacWhinney, 2012). These early learners' L2 categorization patterns should reflect their confidence in L1 naming (i.e., the extent of L1 dominant name agreement) because, in the absence of L2-specific lexical semantic knowledge, inferences about L2 words are based on knowledge of their L1 translation equivalents. Eventually L2 learners become sequential bilinguals, cross-language influence approaches that of simultaneous bilinguals, and they become less native-like in their L1 as L2 lexical semantic proficiency increases under immersive L2 influence (Pavlenko and Malt, 2011; Malt et al., under review). Typicality

ratings also can be construed as a measure of native speakers' confidence about the name of an object, and Pavlenko and Malt (2011) found that Russian–English bilinguals relied on both Russian and English native typicality norms for individual objects when naming these objects in Russian, suggesting that their intuitions about categorization were influenced by the perceived confidence of each language community in an object's category membership.

It is evident that in many instances of simultaneous and sequential bilingualism, the category information provided to bilinguals by the native-speaker communities of each language is variable and yet still bears a significant influence on their production in both languages. With relatively few lexical category stimulus sets normed for native speakers of more than one language and tested on sufficiently advanced bilinguals of both languages, the exact degree and means of this cross-language influence remains to be explored. However, native category norms that represent the full distribution of names produced and thus the degree of name agreement and variation among native speakers may allow an elaborated view of cross-language competition and transfer. The extent to which L1 representations are vulnerable to change may vary as a function of their own entrenchment, with greater native naming agreement representing more robust L1 representations. Conversely, objects named with greater consistency in L2 (high L2 native agreement) could be associated with better learning outcomes as compared to objects for which L2 speakers show little agreement.

THE PRESENT STUDY

In the present study, we aim to disentangle the respective roles of four broad categories of individual language history variables in predicting native-likeness of L2 lexical semantics: L2 environment (non-immersion vs. immersion), age of L2 onset, years of L2 classroom study, and L2 usage pattern [code-switching frequency (CSFreq)]. Collinearity between age and immersion predictors has been shown to cause serious confounds in studies of second language acquisition (see Stevens, 2006 for detailed analysis). Recent ERP studies of individual L2 word processing have identified both positive (Ojima et al., 2011) and negative (Granena and Long, 2012) effects of age while trying to deconfound the effects of age, exposure, and immersion. Previous studies of lexical categorization have also identified confounded relationships between LOR in an L2 environment and age of onset (Malt and Sloman, 2003) and between age of immersion and patterns of language use or dominance (Pavlenko and Malt, 2011).

By measuring several language history variables together, accounting for the earliest L2 exposure (that is, L2 onset before immersion), and using categorization as a more sensitive measure to inter-personal lexical semantic variation, we aim to make better statistical estimates of each variable's effect. We offer a simultaneous measure of four variables based primarily on the self-reports of Chinese–English bilinguals resident in Beijing, China and in Pennsylvania, United States.

We also introduce linguistic community norms for word use in L1 and L2, derived from native speakers of each language, as possible predictors of bilinguals' lexical categorization patterns. The contribution of such norms has rarely been considered in predicting L2 performance (except see Pavlenko and Malt, 2011).

These non-immersed and immersed participants are compared in an L2 (English) lexical categorization task that has proved highly sensitive to variation in lexical semantic mapping for other populations of bilinguals. Based on the simultaneous evaluation of all four language history variables and the linguistic community norms, we evaluate participants' English native-likeness on the lexical categorization task. We offer an interactive account of how various aspects of one's native language, second language, and language learning history jointly influence the lexical semantic mappings that defines object naming, a behavior that occurs often in our daily experience.

MATERIALS AND METHODS

PARTICIPANTS

Two groups of bilingual students, one in the United States and one in China, participated in this study. In the U.S., Chinese–English bilingual undergraduate and graduate students were recruited from the Introduction to Psychology subject pool and through posters around the campus community at Penn State University (State College, PA, USA). In China, Chinese–English bilingual undergraduate and graduate students were recruited through an online campus message board (BBS) and through personal referrals at Beijing Normal University (Beijing, China). Generally speaking, the students at Penn State were slightly younger (mostly undergraduates) than those at Beijing Normal (mostly graduate students), were first exposed to English at a slightly earlier age, and had higher self-rated proficiencies in English.

Although many of the bilingual participants reported some degree of training in a third language, most rated themselves at very low proficiency. Participants who self-reported a proficiency of 2.5 or greater in the third language on a 7-point scale (averaged across four ratings: reading, writing, speaking, listening) or failed to provide a proficiency rating in their third language were not included in the data. In total, 57 participants from Beijing Normal and 68 participants from Penn State met the inclusion criterion. Third languages included French, German, Russian, Mongolian, Japanese, Korean, Taiwanese, and Cantonese.

Penn State students ranged in age from 18 to 23 ($M = 19.5$, $SD = 1.2$). They were first exposed to English between ages 1 and 16 ($M = 8.2$, $SD = 3.7$), and self-rated their English proficiency between 2.5 and 7.0 ($M = 4.7$, $SD = 1.3$). The Penn State students had resided in the United States for 0–19 years ($M = 5.0$, $SD = 5.9$). Students at Beijing Normal University had ages ranging 18 to 28 ($M = 22.8$, $SD = 2.0$), and age of earliest English exposure was 5–15 ($M = 11.4$, $SD = 2.1$). Their self-rated English proficiency varied between 1.3 and 5.5 ($M = 3.9$, $SD = 1.0$), as some were studying English while others majored in different subjects. None of the participants at Beijing Normal University reported living in or visiting an English-speaking country for an extended period of time.

We also drew on a set of native-speaker norming data from functionally monolingual participants who had participated in a previous version of the lexical categorization task, using the same stimuli (Malt et al., 2013). The picture naming data for 25 native

Chinese speakers in China and 28 English speakers in Pennsylvania provided linguistic community norms for the current analyses. Their choices represent the most likely input patterns for bilinguals in their respective language environments.

MATERIALS

All participants completed a language history questionnaire (LHQ; Li et al., 2006) to assess bilingual status, L2 proficiency, age of L2 acquisition, and behavioral predictors such as patterns of code-switching. The LHQ was available in both English and Chinese (simplified characters) and administered according to the dominant language environment. With respect to code-switching, the LHQ allows participants to self-rate their frequency of code-switching in four contexts: Spouse & Family, Friends, Co-Workers, and Classmates. Participants ranked their CSFreq in each context ordinally, using response options that ranged from "Rarely" to "Very Frequently." These responses were transformed into a Likert score between 1 and 5 and averaged within context group to produce the CS scores.

Early trials at Penn State revealed that several participants failed to complete the code-switching section of the LHQ or claimed to never code-switch, a self-report that may (in some cases) underestimate the true rate of code-switching in cultural environments that stigmatize language mixing. An additional code-switching questionnaire (CSQ) was added to subsequent sessions to specifically probe participants' code-switching and was administered according to the dominant language environment. A single item on the CSQ was used to obtain a point-estimate of participants' overall CSFreq: "Do you use English words when speaking Chinese, or do you use Chinese words when speaking English?" rated on a five-point ordinal scale with response options from "never" to "very often."

Sixty-seven photographs of common household objects were used to elicit category names from monolingual and bilingual participants. These objects were drawn from a stimulus set (called the dish set) used by Ameel et al. (2005) to reveal cross-language lexical categorization differences in Dutch–French bilinguals. Each photograph contained a single household serving vessel (e.g., a plate, cup, or bowl) on a neutral background and a centimeter ruler in the foreground for scale (see **Figure 1**). Photographs were displayed at 480 × 360 pixels on a personal computer equipped for digital recording. Each voice response was recorded through a standard omni-directional consumer microphone to the computer's sound card and encoded as 10 s uncompressed WAV files. Each photograph was accompanied by the written prompt: "What is this?" or "这是什么?" according to the task language.

An Operation-Span (O-Span) test was also used to screen the bilingual participant groups for systematic differences in working memory, a cognitive factor that might be confounded with language proficiency or language transfer. The O-Span includes mathematical and verbal components (Turner and Engle, 1989): Participants judge the accuracy of math equations and are provided a word to remember after each judgment. After several math and word combinations, participants are prompted to recall the words they have seen. Arabic numerals were used for the math component (consistent with both Chinese and American math education) and Chinese characters were used for the verbal component. Participants entered their judgments using a computer keyboard and recorded their verbal responses on a paper worksheet. No significant difference was found between the two bilingual samples in their O-Span scores.

PROCEDURE

After giving informed consent in the local language, participants completed the LHQ, also in the local language (Chinese or English). They then completed an unrelated English receptive vocabulary task (results not presented here) to establish an English language mode to the extent possible in both the Chinese- and English-immersed participants. After the vocabulary test, all participants performed the English picture naming task. The Chinese O-Span was then completed and used to shift participants into a Chinese language mode before naming the objects again in Chinese. Finally, the CSQ was completed last. Participants in the US completed English and Chinese tasks on separate days, 1–2 weeks apart (range: 6–21 days; mean: 9 days) and counter-balanced for order. Sessions in China could not be scheduled separately and all tasks were completed on the same day, with English first, followed by Chinese. We reasoned that the English task was less likely to influence Chinese naming in a Chinese immersion environment, and intervening Chinese tasks (namely, the O-Span) would help to reduce any language priming effects.

In the picture naming tasks, participants were instructed to name aloud photographs of objects depicted on the computer. They were asked not to name the objects' contents, as illustrated by two photographic examples: a grocery bag full of vegetables (called *bag*) and a trash can full of paper (called *trash can*). These instructions were provided in written form on the computer screen according to the language of the task. Participants were verbally encouraged to name every object and to always make a guess if unsure. Participants were also provided two practice naming trials for photographs of unrelated bottle-like stimuli, followed by the most dominant monolingual name for each stimulus (*bottle* or *瓶子*) to demonstrate the desired response type. Participants were permitted to take as long as needed to name each picture to ensure that they selected what they considered the best name for each object. Due to disk storage constraints, only the first 10 s (from the onset of the stimulus) of participants' responses were digitally recorded by the computer for each stimulus.

DATA ANALYSIS

Participant responses were transcribed from audio recordings by high-proficiency Chinese–English bilinguals in the United States who were able to comprehend Chinese responses and phonetically accented English responses. Transcribers were not able to view the objects during transcription to prevent bias on ambiguous recordings. Transcribed responses were subsequently reduced to head nouns (e.g., "a small blue bowl" is reduced to "bowl") for comparison with the native norming data. Skipped trials, inaudible responses, and irrelevant responses (e.g., "I don't know") were entered as blanks and treated as missing data.

Four biographical variables were included for each subject: Age of first exposure to English (AOEE), LOR in the English

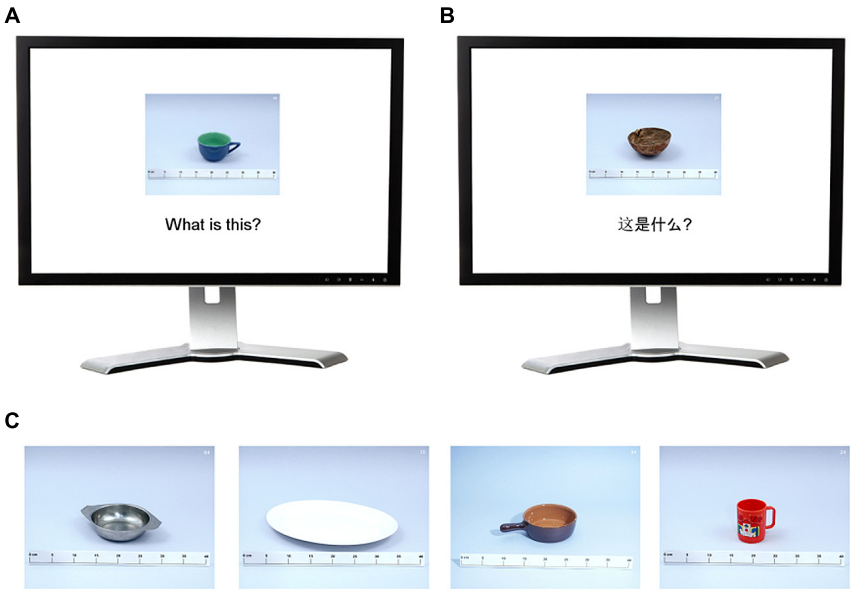


FIGURE 1 | Sample displays for English (A) and Chinese (B) naming tasks depicting two example objects from the dishes stimulus set. (C) Additional example objects. Participants viewed all 67 objects in this format with the text prompt (according to task language) provided below each item.

immersion environment (LOR), self-reported frequency of code-switching between Chinese and English (CSFreq) and the total number of years spent learning English (current age minus the age of first exposure, YrsLearn). For participants who failed to complete some language history and code-switching questions, missing data for the CSFreq variable were replaced with the sample mean (3% of the participants included in the analysis). Participants who did not report AOEE were excluded from the analysis (eight participants), and an additional set of early childhood bilinguals (AOEE $<$ 5 years, six participants) were removed from the US participant data to maintain comparability with the sample in China (AOEE range 5–15 years).

Given the above exclusion/inclusion criteria for data analysis, our data analyses presented in the Results section were based on a total of 30 participants from China and 33 from State College (see **Table 1**). Results from 20 participants in the US sample and 20 in the China sample were discarded due to recording equipment failure (no audio data recorded). Participants were encouraged to speak loudly and clearly directly into a desktop microphone; however, additional participants (not excluded due to recording failure) periodically produced inaudible responses or no response, decreasing their total response rate[1]. Some participants may have chosen not to respond to stimuli when uncertain about those objects' names, a possibility we tested by correlating response rate to self-rated English proficiency. Indeed, response rate in English was weakly correlated to English proficiency ($r = 0.24$, $p = 0.047$) while response rate in Chinese was not ($r = 0.20$, $p = 0.101$). Non-response as a predictor of name uncertainty is preserved in the remaining participants (50% or higher response rate) insofar as all trials are included for analysis, with non-response trials counted as incorrect names.

[1]Two participants in the US and four in China were excluded for response rates below 50% on one or more of the naming tasks. Non-response rates were approximately the same between the English task (four participants in Beijing) and the Chinese task (three participants in Beijing and two in State College), suggesting that most of these missing data were attributable to participant inattention. Two participants in China were removed for naming accuracy scores more than 2.5 standard deviations below the mean (see Subject-wise Analysis).

Table 1 | Demographics and language histories of participants before and after screening.

Sample	*n*	Age (SD)	AOEE (SD)	EngProf (SD)	CSFreq (SD)	LOR (SD)
All participants						
Beijing Normal	57	22.9 (1.8)	11.6 (1.9)	4.1 (1.0)	1.1 (1.2)	0
Penn State	68	20.9 (2.9)	8.8 (3.3)	4.7 (1.1)	1.8 (1.2)	3.8 (5.2)
Included participants						
Beijing Normal	30	22.8 (1.7)	11.5 (2.2)	4.2 (1.1)	0.95 (1.1)	0
Penn State	33	21.8 (3.3)	9.8 (2.6)	4.7 (1.0)	1.9 (0.8)	2.2 (2.6)

RESULTS

In the following sections, we present a set of analyses that examine the lexical categorization patterns of the Chinese–English bilingual participants at three different levels, as follows. (1) The group-wise analysis compares the overall patterns of transfer and convergence between Chinese and English as spoken by the bilingual participants. This analysis looks at the overall trends in naming distributions generated by sub-groups, which is defined by their degrees of L2 immersion (see details below). This analysis allows direct correlations of the bilinguals' overall patterns with the monolingual norms. (2) The subject-wise analysis focuses on individual bilingual participants' language histories and how these variables predict their individual differences in L2 naming patterns. (3) The item-wise analysis examines naming performance on each object of the stimulus set, controlling for variation in individuals' language histories and examining the impact of linguistic community norms on the bilinguals' accuracy in producing the native preferred L2 names.

GROUP-WISE ANALYSIS: CROSS-LANGUAGE TRANSFER AND CONVERGENCE

For group-wise comparison, participants were organized by three discrete values of LOR to describe three types of immersion conditions observed in our sample: No Immersion, Short-term, and Long-term. No Immersion was defined by LOR = 0, describing participants who have never lived in an English immersion environment. English-immersed participants were divided into two groups by a median split (median non-zero LOR = 1.3 years). Short- and Long-term Immersion were defined as the samples below and above the median, respectively.

A cross-language correlation matrix was calculated for each bilingual and monolingual group according to the method of Malt et al. (1999; see also Ameel et al., 2005), in which the naming distribution for each object over all possible names is correlated with the naming distribution for every other object. This method produces a 67 × 67 correlation matrix with 2211 unique values (per speaker group) for our data, indicating, for each possible pair of objects, to what extent the same names were produced with the same frequency by the speaker group. These inter-object matrices can then be correlated between languages or groups, representing the degree to which objects names are distributed similarly in the two samples (regardless of the actual names themselves). **Figure 2** provides these correlation matrices for each immersion group, compared with the monolingual speakers of Chinese and English and between the Chinese and English patterns produced by the bilinguals, according to the convention of Ameel et al. (2005).

The cross-language correlations revealed that native, monolingual speakers of Chinese and English correlate in their categorization of this set of objects at $r = 0.64$ (see the top row of **Figure 2**). This value serves as the baseline correlation against which bilinguals' Chinese and English categorization patterns can be compared. All of the bilinguals showed a highly convergent pattern of naming between languages, correlating their Chinese and English word use around 0.92–0.95 (the bottom row of **Figure 2**), strongly suggesting that they relied on a single set of mappings (with varying degrees of influence from each language). Correlation of the bilinguals' English naming with the monolingual norms (the right-most vertical connection for each matrix in **Figure 2**) was compared using Cohen and Cohen's (1983) method for comparing correlation coefficients. Similarity to the English norms was highest in the Long Immersion group (0.81, compared to 0.78 and 0.77 in the No and Short Immersion groups respectively, $p < 0.01$ in both cases). The bilinguals' English categorization also decreased in its dependence on Chinese norms with increased immersion (No Immersion: 0.80 vs. Short Immersion: 0.78, $p = 0.038$; No Immersion vs. Long Immersion: 0.74, $p < 0.001$).

Surprisingly, the No Immersion group showed the highest convergence between their two languages (0.95), a relatively low correlation with the monolingual Chinese (0.80) compared to their recently immersed peers (Short Immersion, 0.85, $p < 0.001$) and greater Chinese resemblance to the English patterns (0.77, compared to Short Immersion 0.75, $p = 0.058$). This effect may be attributable to differences in the administration of the Chinese naming task, in which the No Immersion group completed Chinese naming shortly after the English naming task. Henceforth, we will examine English naming only, as English names were not subject to priming across tasks because English naming occurred either first or in a separate session for all participants, and L2 acquisition is the focus of the current study.

SUBJECT-WISE ANALYSIS: THE ROLE OF LANGUAGE HISTORY VARIABLES

Participants' picture naming responses were compared to a set of English native norms to generate a score for each participant describing the English native-likeness of their lexical categories.

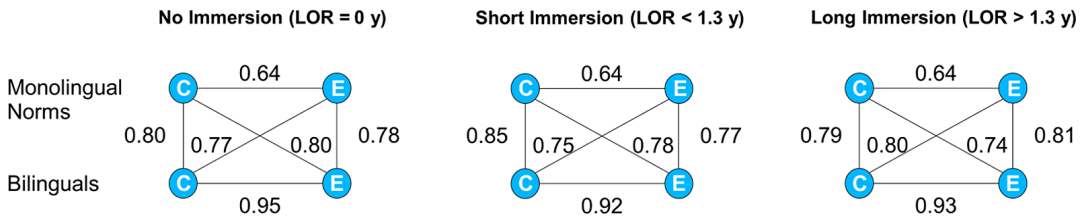


FIGURE 2 | Cross-language correlation matrices for each Immersion group, representing the correlation between lexical categorization patterns in Chinese (C) and English (E) for the bilinguals and native speaker norms.

Each of a participant's responses was awarded a score based on the proportion of native monolingual speakers who produced that same response in the norms (following Malt and Sloman, 2003). Thus if an object was called *mug* by 75% of the norming group and *cup* by 25%, the bilingual participant would receive 0.75 points for naming the object *mug*, 0.25 for *cup*, and 0 for anything else. These point values were averaged across the 67 objects for each participant, rendering an agreement-weighted native-likeness score ranging between 0 and 0.68 (the mean of agreement level for native English speakers across all objects).

We estimated a linear regression model for the English native-likeness scores over participants' language histories to determine the relationships between language background and attained L2 lexical category proficiency. Previous analyses from smaller datasets showed several two-way interactions between the language history variables and an inter-dependency of the significance of these interactions in the model (see Zinszer et al., 2012, 2013 for examples). Consequently, the initial model was estimated with all possible interactions (up to four-way) and, indeed, yielded several highly significant three-way interactions [omnibus test: $F(15,46) = 2.14$, $p = 0.02$, Adjusted $R^2 = 0.22$].

In an attempt to improve the parsimony of this model without discarding important interaction effects, an automatic Akiake information criterion (*AIC*) stepwise procedure was adopted which started with all possible interactions and systematically excluded and re-included variables to find the best-fitting model with the lowest *AIC* score (Venables and Ripley, 2002). This method produces a reduced model while minimizing impact on the model's fitness to the data. Finally, the AIC search excluded only the four-way interaction term, resulting in a significant reduced model [omnibus test: $F(14,47) = 2.27$, $p = 0.02$, Adjusted $R^2 = 0.23$] with slightly improved parsimony (initial model: 16 terms, $AIC = -103.9$; reduced model: 15 terms, $AIC = -105.0$). Ultimately, all predictors were included in one or more significant interactions.

To understand the highly interactive terms of the subject-wise model, we generated several estimated marginal means plots based on the model's predicted English native-likeness scores across a range of values for the two-way interactions between L2 immersion (LOR) and each remaining predictor (while holding other predictors constant at the mean value). These two-way interactions were all highly significant (LOR × YrsLearn: $p = 0.002$; LOR × AOEE: $p = 0.014$; LOR × CSFreq: $p = 0.007$). To further simplify the plots, we again used the three discrete values of LOR to describe three types of immersion conditions observed in our sample: No Immersion (LOR = 0), Short- (LOR < 1.3 years), and Long-term (LOR > 1.3 years). Short and Long-term Immersion were represented by the mean LOR values for each of these two groups: 0.5 and 4.7 years, respectively. **Figure 3** shows plots for the interactions between LOR and each of the remaining predictors: AOEE, CSFreq, and YrsLearn.

The first plot (**Figure 3A**) contrasts years of English study with the duration of English immersion, which are not independent predictors. That is, as the duration of immersion (LOR) increases, so do the years of English study (YrsLearn). Conversely, however, YrsLearn may increase without immersion experience. Therefore these variables contrast the predicted English native-likeness associated with varying durations of study when the amount of that time spent in an Immersion environment is held constant (in this case at 0, 0.5, or 4.7 years). In all three LOR conditions, the relationship between YrsLearn and English native-likeness is negatively sloped, indicating a relative disadvantage for years of English study after controlling for years of English immersion. In other words, every additional year of English study in China beyond a participant's immersion experience reduced the native-likeness of their English categorization patterns. For example, a learner with 15 years of study and almost 5 years of immersion (LOR = 4.7) has had over 10 years of English study in China, and they are predicted to perform worse (on average) than somebody with fewer years of English study and the same amount of (or even less) immersion.

Age of earliest English exposure (AOEE; **Figure 3B**) also displayed a negative relationship with English native-likeness. When controlling for the other variables, later ages of English onset generally result in poorer performance in English categorization. Interestingly, however, the interaction with LOR did not appear to be large (the lines are roughly parallel) indicating a largely additive effect of these two variables. The relative weight of each variable was approximately balanced such that the negative effect of being

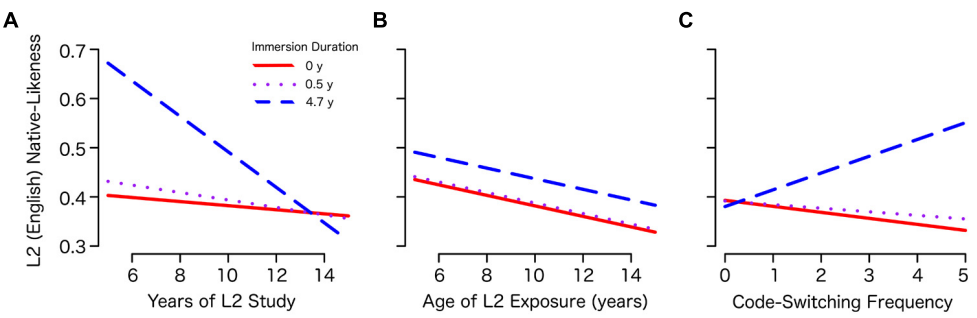


FIGURE 3 | Two-way interaction plots for Immersion and each of the learner variables. (A) Years of L2 study. **(B)** Age of earliest L2 exposure. **(C)** Self-reported frequency of code-switching.

exposed to English 1 year later is offset by the benefit of 1 year of immersion experience.

Participants' self-reported CSFreq was also a significant predictor in the model and significantly interacted with immersion. As **Figure 3C** indicates, the effects of CSFreq were relatively small for non-immersed learners and learners with relatively little immersion experience, and greater CSFreq was associated with less L2 achievement. However, CSFreq was a much stronger predictor for learners with longer immersion experiences, and the direction of the influence was opposite, showing significant gains in English native-likeness with greater frequency of switching between languages. This interaction may suggest that CSFreq is most predictive for people who are immersed in the L2 environment.

ITEM-WISE ANALYSIS: THE IMPACT OF LINGUISTIC COMMUNITY NORMS

In this analysis looking at how native naming consensus for objects impacts the likelihood of naming objects correctly, we compared each response by the participants to the single dominant name[2] produced by the English native norm for each given object. Thus, trials in which the participant produced the norm's dominant name were scored as 1 (correct), while all other trials were scored as 0 (incorrect). Next, we performed two binomial logistic regressions to estimate the probability that a participant would produce the dominant name for any given object.

In the first logistic regression, we entered the same language history variables used in the subject-wise analysis to determine how adequate these variables were for identifying variation in native-like categorization for different objects. The logistic regression model including only participants' language history information contained several statistically significant predictors, but offered a very poor fit to the data (Nagelkerke $R^2 = 0.02$, indicating that subjects' language backgrounds could account for overall trends in the native-likeness of their English categorization but not for most of the variation trial-to-trial. This result points to the importance of considering variation in the learner's input across objects (such as the native norms) as a predictor of success in naming individual objects.

In the next analysis, we added four language variables which described the native speaker norms for every given object: naming agreement in Chinese (L1), naming agreement in English (L2), number of alternative names produced by the Chinese norming group, and number of alternative names produced by the English norming group. Due to computational limitations, this model was estimated with up to four-way interactions and reduced using the same *AIC* stepwise search procedure described in the subject-wise analysis. The resulting reduced model improved the *AIC* compared to the initial model and included 36% fewer terms than the initial model without a serious decrease in fitness (initial model: 163 terms, $AIC = 5036.52$, Nagelkerke $R^2 = 0.25$; reduced model: 104 terms, $AIC = 4951.7$, Nagelkerke $R^2 = 0.24$).

As in the subject-wise analysis, the model contained many interaction terms that impeded interpretation without isolating a few of the variables. Again, we sought to describe how immersion experience affected the role of these language variables in predicting the participants' success in producing native-like English names for objects. A binomial logistic regression predicts the probability that an outcome will occur, in this case the probability that the participant will produce the English native-like dominant name for a given object. Again, we estimated plots in which the individual variables (this time, language variables) interacted with three levels of immersion while holding all other variables constant at a mean value.

Figure 4 presents plots of each linguistic community variable against the three levels of English immersion (None, Short-term, and Long-term). These plots revealed that English native-likeness in the learners was more likely at higher levels of English norm agreement (**Figure 4A**), while the inverse was true for Chinese: There was less English native-likeness with higher agreement in the Chinese norm (4B). An opposing relationship was also observed for the number of alternative names available from the norming sample. Having a greater number of English names available in the norm actually increased the predicted probability that the learners would produce the dominant English name (4C), but having many possible Chinese names for an object decreased the predicted probability of the participants producing an English native-like name (4D). The apparent advantage for a greater number of English names is explored in the next section.

In a follow-up analysis, we asked how L1 and L2 norms might interact with one another in predicting a learner's success in producing the L2 dominant name. Several interaction terms between

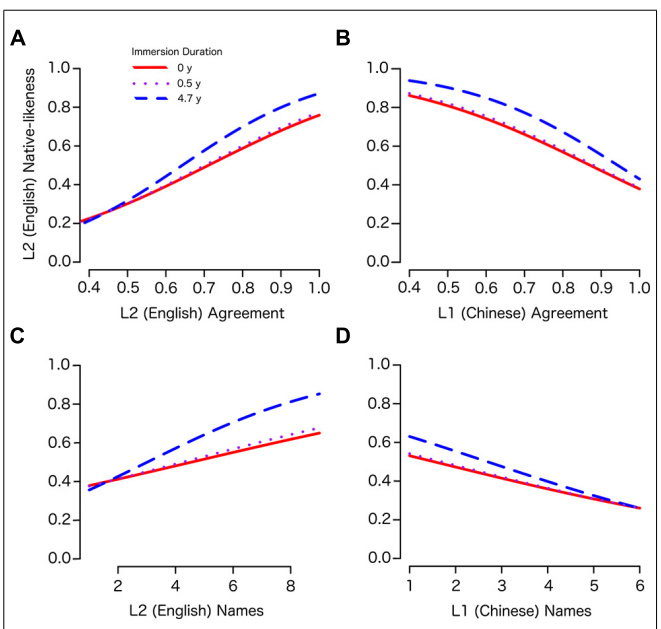


FIGURE 4 | Two-way interaction plots for Immersion and each of the linguistic community norm variables. (A) L2 native speaker agreement (percent of a norming sample who produced the dominant name) **(B)** L1 native speaker agreement. **(C)** Number of alternate names for an object produced by the L2 native speakers in a norming sample. **(D)** Number of alternate names produced by L1 norming sample.

[2]In two cases, the naming agreement score for an object was tied between two names. For each case, we randomly selected one name as the "dominant" name for the purpose of the comparison. This uncertainty, however, is preserved in the L2 Name Agreement variable included in the logistic regression.

these norming variables were highly significant, so we examined the cross-language relationships between L1 and L2 agreement and number of alternate L1 and L2 names. This analysis offers a closer examination of two interesting effects from the preceding results: (1) native speaker agreement in each language appears to compete in predicting L2 native-likeness and (2) an increasing number of names in English seems to be associated with greater L2 native-likeness.

Figure 5 depicts both the observed item-wise accuracies (A and C) and the estimated marginal mean accuracy at varying levels of the predictors using the logistic regression model. In **Figure 5A**, the average response accuracy across participants is plotted over both English and Chinese norm agreement levels for each object in the stimulus set. This plot shows the empirical effects observed in our sample and is generally consistent with the competitive account of L1 and L2 agreement estimated by the regression model. Among the objects in the experimental stimulus set, best performance (depicted by blue-colored dots) is observed when *both* languages have high agreement levels. Items in this set are not limited to one-to-one translation pairs, as *cup*, *mug*, and *glass* were all represented in this set (and all translated as *bēizi*).

This performance diminishes as English agreement decreases. In general, high levels of L2 (English) agreement are associated with successful learning across varying levels of L1 (Chinese) agreement. However, the worst performance by the learners occurs when Chinese agreement is high and English agreement is low, confirming that L1 patterns can have a strong negative effect on L2 native-likeness when L2 input is inconsistent. Many of these items came from the *bēizi* (roughly, *cup*) category, highlighting learners' weakness with the sub-divided English categories it includes (*cup*, *mug*, and *glass*), but several cross-cutting categories that did not have clear one-to-one translations also appeared: *gāng*, *pénzi*, and *wǎn* are approximately translated as decreasing sizes of *bowls* but are translated as *dish* for some items according to the monolingual norms.

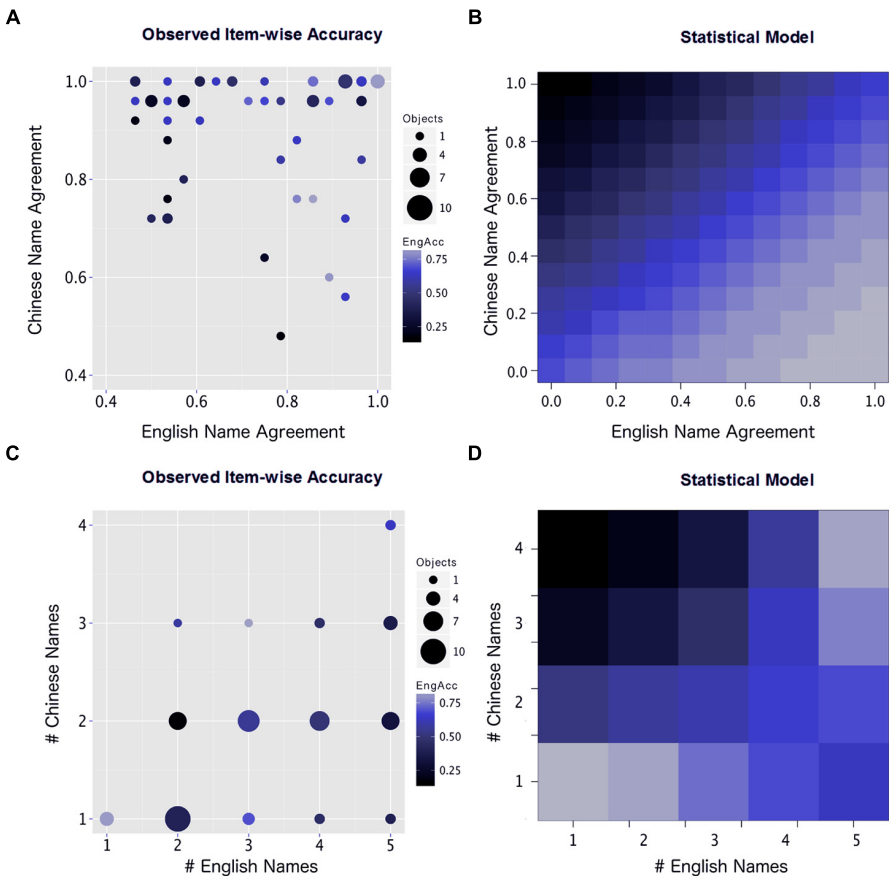


FIGURE 5 | Observed and estimated effects of language-specific variables: dominant name agreement and number of alternate names. **(A)** The interaction between native speaker naming agreement in Chinese and English as predictors of English naming accuracy by Chinese–English bilinguals. Each point on this graph represents a set of objects with particular agreement values in Chinese and English. Color of each point indicates mean accuracy ratings for the participants, and size of each point indicates the number of objects represented. **(B)** The accuracy values estimated by the statistical model at varying levels of Chinese and English agreement, controlling for all learner variables and number of alternate names. **(C)** The interaction between the number of names generated by native speakers of Chinese and English as predictors of English naming accuracy by the Chinese–English bilinguals. **(D)** The accuracy values estimated by the statistical model at varying numbers of alternative names in Chinese and English, controlling for all learner variables and the level of native speaker agreement.

For comparison, **Figure 5B** plots the model's estimated accuracy levels, generalized to all levels of agreement in each language. While **Figure 5B** covers a broader range of potential values (such as low naming agreement in Chinese, which is under-represented in the actual stimulus set), it generally fits the patterns established by the empirical data. Further, the observed data (**Figure 5A**) do not control for confounding variables (such as the number of alternate names). The regression model estimates both predictors and thus isolates the effect of agreement while holding the number of names constant, resulting in the smoother contours along values of L1 and L2 agreement depicted in **Figure 5B**.

Participants' observed performance across the different numbers of alternative names in the English and Chinese norms, however, differed significantly from the regression model's estimated accuracy rates. **Figure 5C** depicts the accuracy rate for each object in the stimulus set across varying numbers of alternate names in each language. This plot indicates that when learners had only one L1 name for an object in each language, performance was highest. The worst performance was observed when exactly two competing names for an object were available in either language. The model's prediction that more competing L2 names improve the probability of learners producing the L2 dominant name (**Figure 5D**) is consistent with the latter observation that objects with three or more competing names were named more accurately than those with two names, but it overlooks the advantage for objects with only one name. Again, in the observed data (**Figure 5C**), the number of names and agreement level are confounded, but the regression model isolates these effects and controls for agreement in its estimations of the effects of L1 and L2 names (**Figure 5D**). It is not clear that these representations disagree, *per se*. Rather, **Figure 5C** represents the objects provided in the stimulus set, while **Figure 5D** provides a controlled, parametric representation over many values of each variable. The discrepancy between these representations is further discussed below.

DISCUSSION

SUMMARY

In this study we examined the relative effects of four language history variables in predicting learners' outcomes in L2 lexical categorization native-likeness. Highly significant interactions were found among these variables, supporting the idea that language history (e.g., age of L2 onset) variables should not be evaluated in isolation from other variables. Significant age of L2 onset effects were observed, but these effects were tempered by the positive contribution of increased immersion experience. A surprising observation was that increased experience with L2 prior to immersion was actually associated with reduced native-likeness of L2 lexical categorization. Finally, we found that for bilinguals with long-term L2 immersion, patterns of language use (i.e., code-switching habits) were a significant predictor of L2 native-likeness, but for learners with less immersion experience (including no immersion experience), language use was a less important predictor of L2 native-likeness.

We further explored how the naming norms of the linguistic communities of both languages influenced the learners' success in acquiring native-like L2 lexical semantic mappings. Both L1 (Chinese) and L2 (English) norms were significant predictors of the learners' L2 native-likeness, consistent with previous findings in other domains of second language acquisition, such as phonology. Further, we identified unique effects for agreement among native speakers and the number of alternate names produced in the norming samples. The result of an item-wise analysis revealed that a large amount of the between-object variation in naming was captured by these native speaker naming norms, indicating both the lasting impact of L1 mappings on L2 production and the sensitivity of L2 learners to the native speaker norms of the L2. Below we present a more detailed discussion of how L2 naming patterns are influenced by the learner variables and input (linguistic community norm) variables.

LEARNER VARIABLES

L2 training

The most surprising finding of this study was that the number of years spent studying English outside an immersion environment was negatively related to L2 native-likeness in the lexical categorization task, even after controlling for the length of eventual L2 immersion. This outcome was not predicted by any past research nor intuition. This novel contrast between years of non-immersed and immersed learning in learners who have significant experience in both environments suggests that L2 training outside of an immersion environment may ultimately reinforce lexical semantic mappings that significantly differ from those of L2 native speakers. There is little doubt that immersion experience is beneficial to second language learning, and second language acquisition research has long promoted this view, but the present study adds the unique corollary that L2 learning without immersion may, in fact, hinder native-likeness. This effect may be due to the entrenchment of L1 structures in learners' L2 as a result of impoverished input. Common classroom techniques for learning translation equivalents or naming highly prototypical objects encourage learners to export their inferences about object categories from L1 to L2 by way of one-to-one translation. However, native-like L2 mappings only become available to the learner with more diverse input from an immersion environment or (potentially) another immersive instructional setting such as the highly enriched virtual environments that may be simulated in computer games (see Legault et al., 2014, for example). The more time that L2 learners spend learning lexical semantic mappings in a non-immersive environment, the more entrenched the L1-driven mappings in L2 may become. Considered against this perspective and the relative proportion of L1 vs. L2 input in the non-immersive environment, the patterns in our data become less surprising but provide an important lesson for language instruction practice.

Age of onset

Second language lexical learning has often been regarded as a qualitatively different type of acquisition from phonology and syntax that tend to show strong age effects. One theoretical account, Ullman's (2001) Declarative-Procedural (D-P) Model, attributes this dissociation to differences in the underlying memory systems that support lexical learning and all other aspects of language. This theory is consistent with observations to date about both native

and second language acquisition, but the present findings suggest that the dissociation may not be so clear cut. When we measure lexical semantics as a complex system of mappings for making generalizations rather than just a set of word-object pairs, as in the present study, a weak pattern of age effects is replicated.

Age of second language onset effects may also be confounded with the negative pre-immersion learning effect. In the present study, we surveyed participants' earliest exposure to English as a second language rather than their earliest immersion experience in an English language community. Although there were significant advantages for earlier learners over later learners, these advantages were limited in the sense that for every year of earlier acquisition, the same effects could be gained by an additional year of L2 immersion. With a small age of onset advantage on the one hand, and a non-immersed L2 learning disadvantage on the other hand, one may ask whether earlier L2 instruction is indeed beneficial for lexical semantic native-likeness. Addressing this question requires considering the multiple influence of both age effects, amount of total training, and the eventual onset of immersion (if at all). In a later section on implications for L2 instruction, we address these issues in further detail.

Code-switching frequency

Whereas age effects and training effects focus specifically on the conditions under which learners begin acquiring a new language, eventual native-likeness may just as well depend on how that language is used at later stages, such as in an L2 immersion environment. Switching from one language to another may be common, even difficult to avoid, in bilingual environments, but considerably more variation in individual CSFreq could be observed among bilinguals in relatively monolingual-like environments. While some bilinguals may use each language in a distinct context (e.g., home vs. work), others may switch frequently. Research in first language lexical attrition has highlighted the role of bilinguals' specific language use patterns in re-shaping L1 (De Leeuw et al., 2010) and offered a cognitive explanation for how L2 structures are eventually encoded into L1 representations (Wolff and Ventura, 2009).

In the present study we observe a complementary effect. Increased code-switching is associated with greater L2 native-likeness. However, this effect interacts with L2 immersion such that it applies only after a significant period of immersion (illustrated at 4.7 years in **Figure 2**). For learners with significantly less immersion (including no immersion at all) code-switching behavior had no strong effect on L2 native-likeness, both emphasizing the importance of prolonged L2 exposure for the acquisition of these lexical semantic mappings and perhaps mitigating a belief that frequent switching between languages significantly impedes native-like acquisition of an L2.

The causal relationship between CSFreq and native-likeness cannot be determined from our results, however. One explanation would argue that increasing an advanced learner's code-switching leads to improvements in L2 native-likeness by promoting simultaneous activation and therefore increasing opportunities for lexical semantic remapping. On the other hand, bilinguals with greater L2 native-likeness may already be more involved in bilingual social settings (as opposed to seeking out L1 contexts) and increase their rate of code-switching as a result. Future research could investigate the short-term effects of code-switching in an experimental procedure, but the long-term causal relationship between these variables remains unknown.

LANGUAGE VARIABLES

Although the learner-oriented variables as discussed above proved useful in predicting overall performance in lexical categorization, they were rather inadequate in predicting native-like naming for individual objects. Language-specific variation, on the other hand, proved extremely important in predicting trial-by-trial accuracy of participants' object naming, even after controlling for inter-participant differences in the learner variables. These effects have been revealed by our item-wise analyses. One lesson from these effects is that any kind of overall attainment score in lexical categorization masks significant variation in mastery for individual words, with some words posing much greater challenges for the learner (see also Malt and Sloman, 2003). The current data help reveal the source of the variation.

Native speaker agreement

We found a competing relationship between the level of native-speaker agreement in L1 and L2 in predicting the native-likeness of learners' L2 responses. The role of L2 agreement in learners' responses indicates that these learners are sensitive to variation in native speakers' lexical categories for these objects. In the alternative case, where learners rely only on a general majority name for objects, we should see little effect of the L2 agreement variable, as learners would be more consistent than native speakers. Instead, learners respond proportionally to native speakers in their level of naming agreement. Further, the interaction between immersion and L2 agreement demonstrates that the advantage for high L2 agreement increased with greater immersion: These objects show greater improvement than low L2 agreement objects, which did not improve much even with almost 5 years of immersion.

Conversely, agreement among native-speakers of the L1 significantly impeded native-like naming in the L2, indicating that L1 learners were more resistant to revising their lexical semantic mappings in L2 when L1 native speakers were more consistent, likely showing a higher degree of confidence about the object's category membership. The interaction between immersion and L1 agreement demonstrates that this L1 disadvantage predicts learners' improvement with greater immersion experience. Low L1 agreement words significantly increase in their native-likeness with longer periods of immersion, while higher L1 agreement words show less improvement, highlighting these lexical semantic mappings' resistance to restructuring.

Re-examining the observed accuracy rates for the learners across both L1 and L2 agreement levels, we found an antagonistic interaction between these variables. When L1 agreement was especially high, learners struggled to produce native-like L2 names, even at relatively high levels of L2 agreement. However, when L1 agreement was relatively low, L2 agreement was a better predictor of L2 learners' native-likeness, apparently becoming more salient in the absence of strong L1 cues. The statistical model did not

find such a strong interaction, instead identifying the same opposing main effects of L1 and L2 agreement but without an effect of the very small (though significant) interaction term. It remains to be seen whether the observed interaction is a byproduct of the objects in our particular task or whether the model simply underestimates the importance of this interaction. In either case, the important roles of L1 and L2 agreement norms are apparent, either independently or interactively.

Alternate names

The number of alternate names for an object produced by native speakers in L1 and L2 were also significant predictors of L2 learners' native-likeness and were highly interactive with one another. If learners have only one name for an object in their native language, the model indicated that they would be equally likely to produce the dominant L2 name, regardless of alternatives. However, the observed naming behavior indicated that this trend overlooked a significant variation from L2 norms in the learners' naming. For this subset of objects with only one name in L1, the lowest probability of producing a native-like L2 name occurred when the L2 provided two name alternatives, with greater L2 native-likeness occurring when only one L2 name was available or when three or more L2 names were available. This pattern suggests two mechanisms: (1) the attraction of the 1-to-1 translation, as learners struggle with competing pairs of L2 names, and (2) the competition within a distribution of L2 names, as learners' performance improves with a greater number of name alternatives, showing some indifference to the L2 alternatives when there are several.

In the remaining conditions, when learners have multiple L1 words for an object, a greater number of L2 names appears to offer an advantage in selecting the dominant name. One potential explanation for this effect is the proportion of input that each alternative name comprises for L2 learners. Because the present model looks at both agreement in the dominant name and the number of alternatives, the latter provides an indirect measure of the native-speaker agreement levels for each alternate (non-dominant) name. As the number of alternative names increases, the remaining portion of the norm is divided into smaller parts relative to the dominant name, and thus each alternative name becomes a less salient competitor.

Under the foregoing explanation, we would expect the lowest L2 learner performance to occur when naming agreement is low *and* split with only one alternate name which shares all of the remaining native speaker agreement, e.g., 60% of native speakers name an object *plate* while 40% call the object *dish*. As new name alternatives are introduced, the second- through nth-most dominant names fall off in agreement, e.g., 60% of native speakers call an object *plate*, while 20% call the object *dish*, and 20% call it *platter*. This account of agreement and alternate names emphasizes the competition between names, and successful native-like naming is supported by greater agreement in the dominant name *relative* to the alternate names available. In our sample of objects, naming agreement and number of alternate names are correlated such that higher agreement is associated with fewer alternate names, producing the low performance effect at two alternate names (**Figures 5A,C**). The logistic model, on the other hand, dissociates these two variables and thus does not show this effect in either variable independently (**Figures 5B,D**).

As the number of names in the sample of L1 native speakers increased, L2 learners' native-likeness declined, suggesting that the relative frequency of the dominant name was less important for L1 than the full array of available names. This observation makes intuitive sense, as we would expect the learners to have a more stable, entrenched knowledge of their native language. In the case of L1, participants may simply be sensitive to the presence of names regardless of agreement level, or alternate names in L1 may reflect a more general uncertainty about the identity of an object and, apart from language, its membership in semantic categories with other objects. If the function of an entirely novel object is unknown, even native speakers will have a difficult time settling on the best name for that object because lexical categorization does not strictly adhere to similarity of physical features like size and shape.

Finally, performance on six objects that had two competing names in both L1 and L2 was observed to be the worst overall among the stimuli. This effect is not replicated in the modeled plots because, again, it depends not only on the number of names but on the combination of name agreement and number of names, while the model parametrically varies each of these factors. Indeed, the item-wise observations are consistent with the proposal that the distribution of naming agreement between the two objects in L2 drives the general disadvantage for two-named objects.

RELATIONSHIP TO PREVIOUS MODELS

In the introduction, we explored how theories of lexical and semantic representation could be extended to understanding patterns of lexical categorization. The present study does not directly implement any specific theoretical model, as we observe only the naturally occurring shifts in lexical categorization by Chinese–English bilinguals over their varying language learning and language immersion experiences. Nonetheless, connectionist theories such as the Distributed Feature Model (Van Hell and De Groot, 1998) and computational models (e.g., McClelland and Rogers, 2003) present a useful formalization for how word-feature mappings may be represented and adjusted with simple associative training paradigms (e.g., pairing the word *bottle* with feature sets describing its exemplars).

Specifically, important factors in connectionist training paradigms, such as amount, frequency, and consistency of input are readily translated into the lexical categorization terms used in this study. We quantify the amount of L2 experience (LOR), frequency of the dominant name relative to other names (naming agreement), and alternate names, finding compelling parallels between the associative learning principles that underlie connectionist models and the estimated effects of these variables on L2 categorization. For example, the (weak) age of onset effect observed in the present study concurs with entrenchment accounts of age effects in models of lexical acquisition (e.g., Li et al., 2007; Zhao and Li, 2010).

Entrenchment also provides some explanation for the relative disadvantage in re-mapping L2 categories for objects with high L1 agreement, as high agreement confers greater training frequency

for the dominant L1 name (for a given object presentation). The role of L2 linguistic community variables in predicting learners' native-likeness confirms that learners are sensitive to the relative frequency of several alternate names, showing improved performance when the agreement for the dominant name increases and decreased performance when alternate name competitors increase in frequency (e.g., two names distributed 60–40% versus three names distributed 60–20–20%). We also found support for the interactive relationship between L1 and L2 mappings, as suggested by the models proposed by Van Hell and De Groot (1998) and Dong et al. (2005). Future research may test whether or not manipulating these training parameters produces analogous results in computational simulations of category learning, validating the comparison between lexical categorization in language and lexical semantic learning in connectionist models.

IMPLICATIONS FOR SECOND LANGUAGE INSTRUCTION

The present study offers several new insights into the role of language history, language training, and language use in second language lexical semantic learning. Most importantly, we find that greater time spent studying a second language before immersion predicted lower levels of eventual L2 native-likeness, likely due to the entrenchment of L1-like lexical-semantic mappings. Although we do find an age of onset effect, even after controlling for immersion and duration of language training, the magnitude of this age effect is proportional to the benefits of immersion, and the benefits to L2 native-likeness from early age of onset are small relative to the effects of more pre-immersion training.

On the extreme end, one might propose that pre-immersion language instruction is actually counter-productive to native-like lexical semantic development, and second language education would be best postponed until immersion opportunities arise. However, this viewpoint is impractical for most non-immigrant learners, and likely over-stated, as our analysis of language-specific variables (native-speaker agreement and alternative names) show that learners are, in fact, highly sensitive to the inconsistent input that describes native-like lexical categorization. Lexical semantic learning in non-immersion environments might therefore be improved by introducing learners to a greater variety of referents and the naturally diverse naming patterns associated with those referents, allowing them to develop more native-like intuitions about the relationships between objects that define lexical categories. The method of using a diverse set of naming patterns in second language instruction clearly contradicts the traditional classroom teaching method, in which training focuses primarily on one-to-one translations; such a focus underestimates cross-languages differences, and by our findings, encourages the use L1 patterns for L2 words and therefore impedes learners' later ability to acquire native-like lexical semantic mappings.

ACKNOWLEDGMENTS

We wish to thank Anqi Li for supervising data collection and coding, as well as Peiyao Chen, Patrick Clark, Anqi Li, Jessica Wen, Han Wu, Zhichao Xia, Tianyang Zhang, Dan Zhong, and Lijuan Zou for assistance with participant recruitment, data collection, and language consulting. We also thank Hua Shu at Beijing Normal University for providing lab facilities and equipment for this study. This research was supported by National Science Foundation grants (BCS-1057855; OISE-0968369).

REFERENCES

Ameel, E., Malt, B. C., and Storms, G. (2008). Object naming and later lexical development: from baby bottle to beer bottle. *J. Mem. Lang.* 58, 262–285. doi: 10.1016/j.jml.2007.01.006

Ameel, E., Malt, B. C., Storms, G., and Van Assche, F. (2009). Semantic convergence in the bilingual lexicon. *J. Mem. Lang.* 60, 270–290. doi: 10.1016/j.jml.2008.10.001

Ameel, E., Storms, G., Malt, B. C., and Sloman, S. A. (2005). How bilinguals solve the naming problem. *J. Mem. Lang.* 53, 60–80. doi: 10.1016/j.jml.2005.02.004

Bowerman, M., and Levinson, S. C. (eds). (2001). *Language Acquisition and Conceptual Development.* Cambridge: Cambridge University Press. doi: 10.1006/jmla.2000.2750

Cohen, J., and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Hillsdale, NJ: Erlbaum.

De Leeuw, E., Schmid, M. S., and Mennen, I. (2010). The effects of contact on native language pronunciation in an L2 migrant setting. *Biling. Lang. Cogn.* 13, 33–40. doi: 10.1017/S1366728909990289

Dong, Y., Gui, S., and MacWhinney, B. (2005). Shared and separate meanings in the bilingual mental lexicon. *Bilingualism* 8, 221. doi: 10.1017/S1366728905002270

Fang, S.-Y., Malt, B. C., Ameel, E., and Li, P. (2013). A computational model of semantic convergence in bilingual lexicon. *Talk Presented at the 43rd Annual Meeting of the Society for Computers in Psychology (SCiP)*, Toronto, ON. doi: 10.3389/fpsyg.2010.00221

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *J. Phon.* 15, 47–65.

Forcato, C., Burgos, V. L., Argibay, P. F., Molina, V. A., Pedreira, M. E., and Maldonado, H. (2007). Reconsolidation of declarative memory in humans. *Learn. Mem.* 14, 295–303. doi: 10.1101/lm.486107

Graham, C. R., and Belnap, R. K. (1986). The acquisition of lexical boundaries in English by native speakers of Spanish. *IRAL Intl. Rev. Appl. Linguist. Lang. Teach.* 24, 275–286. doi: 10.1515/iral.1986.24.1-4.275

Granena, G., and Long, M. H. (2012). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Sec. Lang. Res.* 29, 311–343. doi: 10.1177/0267658312461497

Green, D. W., and Abutalebi, J. (2013). Language control in bilinguals: the adaptive control hypothesis. *J. Cogn. Psychol.* 25, 515–530. doi: 10.1080/20445911.2013.796377

Johnson, J., and Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cogn. Psychol.* 21, 60–99. doi: 10.1016/0010-0285(89)90003-0

Kroll, J. F., Bobb, S. C., and Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: arguments against a fixed locus of language selection in bilingual speech. *Biling. Lang. Cogn.* 9, 119. doi: 10.1017/S1366728906002483

Kroll, J. F., Dussias, P. E., Bogulski, C. A., and Valdes- Kroff, J. (2012). "Juggling two languages in one mind: what bilinguals tell us about language processing and its consequences for cognition," in *The Psychology of Learning and Motivation,* Vol. 56, ed. B. Ross (San Diego, CA: Academic Press), 229–262.

Landar, H. J., Ervin, S. M., and Horowitz, A. E. (1960). Navaho color categories. *Language* (*Baltim*) 36, 368–382.

Legault, J., Fang, S., Wang, S., Lan, Y., and Li, P. (2014). Functional and anatomical changes as a function of second language learning. *Poster Presented at the Annual Conference of the Society for the Neurobiology of Language (SNL 2014)*, Amsterdam.

Li, P. (ed.). (2013). Computational modeling of bilingualism [Special issue]. *Biling. Lang. Cogn.* 16, 241–245. doi: 10.1017/S1366728913000059

Li, P. (2014). "Bilingualism as a dynamic process," in *Handbook of Language Emergence,* eds B. MacWhinney and W. O'Grady (Boston, MA: John Wiley & Sons Inc.), 511–536.

Li, P., Sepanski, S., and Zhao, X. (2006). Language history questionnaire: a web-based interface for bilingual research. *Behav. Res. Methods* 38, 202–210. doi: 10.3758/BF03192770

Li, P., and Zhao, X. (2013). Connectionist models of second language acquisition. *Contemp. Approaches Second Lang. Acquis.* 9, 177–198. doi: 10.1075/aals.9.12ch9

Li, P., Zhao, X., and MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cogn. Sci.* 31, 581–612. doi: 10.1080/15326900701399905

MacWhinney, B. (2012). "The logic of the Unified Model," in *The Routledge Handbook of Second Language Acquisition,* eds S. Gass and A. Mackey, (New York: Routledge), 211–227.

Malt, B. C., Li, P., Ameel, E., and Zhu, J. (2013). "Language dominance modulates cross-language lexical interaction," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society).

Malt, B. C., and Majid, A. (2013). How thought is mapped into words. *Wiley Interdiscip. Rev. Cogn. Sci.* 4, 583–597. doi: 10.1002/wcs.1251

Malt, B. C., and Sloman, S. A. (2003). Linguistic diversity and object naming by non-native speakers of English. *Biling. Lang. Cogn.* 6, 47–67. doi: 10.1017/S1366728903001020

Malt, B. C., Sloman, S. A., and Gennari, S. P. (2003). Universality and language specificity in object naming. *J. Mem. Lang.* 49, 20–42. doi: 10.1016/S0749-596X(03)00021-4

Malt, B. C., Sloman, S. A., Gennari, S. P., Shi, M., and Wang, Y. (1999). Knowing versus naming: similarity and the linguistic categorization of artifacts. *J. Mem. Lang.* 40, 230–262. doi: 10.1006/jmla.1998.2593

McClelland, J. L., and Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* 4, 310–322. doi: 10.1038/nrn1076

Ojima, S., Matsuba-Kurita, H., Nakamura, N., Hoshino, T., and Hagiwara, H. (2011). Age and amount of exposure to a foreign language during childhood: behavioral and ERP data on the semantic comprehension of spoken English by Japanese children. *Neurosci. Res.* 70, 197–205. doi: 10.1016/j.neures.2011.01.018

Pavlenko, A., and Malt, B. C. (2011). Kitchen russian: cross-linguistic differences and first-language object naming by Russian–English bilinguals. *Biling. Lang. Cogn.* 14, 19–45. doi: 10.1017/S136672891000026X

Snow, C. E., and Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: evidence from second language learning. *Child Dev.* 49, 1114–1128. doi: 10.2307/1128751

Stevens, G. (2006). The age-length-onset problem in research on second language acquisition among immigrants. *Lang. Learn.* 56, 671–692. doi: 10.1111/j.1467-9922.2006.00392.x

Turner, M. L., and Engle, R. W. (1989). Is working memory capacity task dependent? *J. Mem. Lang.* 28, 127–154. doi: 10.1016/0749-596X(89)90040-5

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Biling. Lang. Cogn.* 4, 102–122. doi: 10.1017/S1366728901000220

Van Hell, J. G., and De Groot, A. M. B. (1998). Conceptual representation in bilingual memory: effects of concreteness and cognate status in word association. *Biling. Lang. Cogn.* 1, 193–211. doi: 10.1017/S1366728998000352

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Chicago: Springer. doi: 10.1007/978-0-387-21706-2

Weber-Fox, C. M., and Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *J. Cogn. Neurosci.* 8, 232–256. doi: 10.1162/jocn.1996.8.3.231

Wolff, P., and Ventura, T. (2009). When russians learn English: how the semantics of causation may change. *Biling. Lang. Cogn.* 12, 153–176. doi: 10.1017/S1366728909004040

Zhao, X., and Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *Intl. J. Biling. Educ. Biling.* 13, 505–524. doi: 10.1080/13670050.2010.488284

Zinszer, B. D., and Li, P. (2010). "A SOM model of first language lexical attrition," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds S. Ohlsson and R. Catrambone (Austin, TX: Cognitive Science Society).

Zinszer, B. D., Malt, B. C., Ameel, E., and Li, P. (2011). Bilingual categorization behavior: internally or externally emergent? *Talk Presented at the 8th International Symposium of Bilingualism,* Oslo, Norway.

Zinszer, B. D., Malt, B. C., Ameel, E., and Li, P. (2012). Predictors of native-like categorization in Chinese learners of English. *Talk Presented at Second Language Research Forum*, Pittsburgh, PA.

Zinszer, B. D., Malt, B., Ameel, E., and Li, P. (2013). What is a cup? Effects of immersion, language, and learner variables on lexical category convergence. *Poster Presented at Psychonomic Society 54th Annual Meeting*, Toronto, ON.

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 May 2014; accepted: 05 October 2014; published online: 27 October 2014.
Citation: Zinszer BD, Malt BC, Ameel E and Li P (2014) Native-likeness in second language lexical categorization reflects individual language history and linguistic community norms. Front. Psychol. ***5****:1203. doi: 10.3389/fpsyg.2014.01203*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 17 April 2014
doi: 10.3389/fpsyg.2014.00329

Naturalistic acquisition in an early language classroom

Anne Dahl* and Mila D. Vulchanova

Department of Language and Literature, NTNU Norwegian University of Science and Technology, Trondheim, Norway

Edited by:
Vicky Chondrogianni, Bangor University, UK

Reviewed by:
Stanka A. Fitneva, Queen's University, Canada
Teresa Cadierno, University of Southern Denmark, Denmark

****Correspondence:***
Anne Dahl, Department of Language and Literature, NTNU Norwegian University of Science and Technology, Edvard Bulls veg 1, 7491 Trondheim, Norway
e-mail: anne.j.dahl@ntnu.no

This study investigated whether it is possible to provide naturalistic second language acquisition (SLA) of vocabulary for young learners in a classroom situation without resorting to a classical immersion approach. Participants were 60 first-grade pupils in two Norwegian elementary schools in their first year. The control group followed regular instruction as prescribed by the school curriculum, while the experimental group received increased naturalistic target language input. This entailed extensive use of English by the teacher during English classes, and also during morning meetings and for simple instructions and classroom management throughout the day. Our hypothesis was that it is possible to facilitate naturalistic acquisition through better quality target language exposure within a normal curriculum. The students' English vocabulary knowledge was measured using the Peabody Picture Vocabulary Test, version 4 (PPVT-IV, Dunn and Dunn, 2007a), at the beginning and the end of the first year of school. Findings are that (1) early-start second-language (L2) programs in school do not in themselves guarantee vocabulary development in the first year, (2) a focus on increased exposure to the L2 can lead to a significant increase in receptive vocabulary comprehension in the course of only 8 months, and (3) even with relatively modest input, learners in such an early-start L2 program can display vocabulary acquisition comparable in some respects to that of younger native children matched on vocabulary size. The overall conclusion is that naturalistic vocabulary acquisition is in fact possible in a classroom setting.

Keywords: second language acquisition, naturalistic acquisition, input, early-start, classroom

INTRODUCTION

Over the past decades, there has been a trend in many countries of lowering starting ages for learning foreign languages, especially English. One reason is globalization and the role of English as an international lingua franca; another is increased knowledge of the benefits of young starting ages for language acquisition. However, the relationship between what we know about language acquisition and what goes on in early language classrooms is not straightforward, and it is not obvious that such classrooms make the best possible use of the learners' young age. A number of studies (e.g., Burstall, 1975; Holmstrand, 1982; Cenoz, 2003; García Lecumberri and Gallardo, 2003; García Mayo, 2003; Lasagabaster and Doiz, 2003; Muñoz, 2006) indicate that an early start in a foreign language does not necessarily make a difference in terms of the pupils' attained competence.

Even though the common assumption that children always acquire languages more easily than adults has been contested (see e.g., Singleton and Ryan, 2004 p. 72 ff. for an overview), the conclusion from findings in research is generally that the earlier one starts acquiring a language before adulthood, the better the chances are of attaining target competence (Johnson and Newport, 1989; Hyltenstam, 1992; DeKeyser, 2000; Hyltenstam and Abrahamsson, 2003; Singleton and Ryan, 2004 ch. 7). This is often attributed to a difference in learning style, as well as maturational constraints related to a sensitive period in language learning (Felix, 1985; Bley-Vroman, 1989; Newport, 1990). Yet little is known about how the factors known to impact on language acquisition interact in the course of development, and what their relative weighting is.

Nikolov (2009) hypothesizes that a possible explanation for the lack of an early-start advantage in previous studies may be that classroom activities employed in that research were better suited to older learners. Quite often the maturational facts in language acquisition link naturally to the learning style differences, namely that younger learners are more likely to employ implicit learning, whereas older learners outperform them on explicit learning (Muñoz, 2006). It then follows that what younger learners need above and beyond all else is exposure to the target language—not explicit instruction and formal training. We know that L2 learners are fully capable of acquiring linguistic knowledge without intentional effort or instruction, and that reading and listening alone can lead to acquisition especially in young learners (cf. Lightbown, 2000; Lightbown et al., 2002). Amount and quality of input are undoubtedly crucial factors in SLA (cf. Hyltenstam, 1992; Gass, 2003), and there is evidence that sensitivity to frequency is relevant for the acquisition of grammatical items (cf. Larsen-Freeman, 1975; Goldschneider and DeKeyser, 2001). Frequency of language items and volume of language exposure have also been demonstrated to influence vocabulary size, at least in L1 acquisition (Hart and Risley, 1995; Childers and Tomasello, 2002; Hoff and Naigles, 2002; Vulchanova et al., 2012), e.g., contributing to learning from distributional cues, a mechanism found in both L1 and L2 acquisition (e.g., Saffran et al., 1996, 1997; Pelucchi et al., 2009).

As Wode (1981, p. 302) points out, "[t]here is no learner on record who learned a language or even part of it without receiving some language input." Instruction and explicit knowledge may play a role in SLA, specifically in compensating for the limited time and opportunity for exposure in the language classroom (cf. Lightbown, 2000). However, it is likely that explicit instruction is less relevant for young learners, and that cognitive maturity may be necessary in order for explicit forms of instruction to make up for impoverished input (see e.g., DeKeyser, 2000; Muñoz, 2001; DeKeyser and Larson-Hall, 2005; Larson-Hall, 2008). There is thus reason to believe that high-input child SLA contexts are the successful ones, and that both intensity and continuation of exposure are decisive factors (Burstall, 1975; Stern, 1983; Lightbown, 2000; Abello-Contesse et al., 2006; Ruiz-González, 2006; Larson-Hall, 2008: 56).

The crucial question is whether acquisition in early-start L2 classrooms can be significantly improved even with only a limited increase in the amount and density of exposure to English. This can be achieved by giving the language itself a more central place in the English classroom, e.g., in conducting classroom management in English and in prioritizing input-heavy activities such as the teacher reading aloud. In addition, L2 input can be increased also outside of English class by providing classroom management and simple instructions in English throughout the school day. The present study is, to our knowledge, the first study to use such an approach, and to investigate the effect of such increased input on vocabulary acquisition in the context of English as an L2 in Norway.

Norwegian children start learning English systematically in school from age 6. However, the number of teaching hours is low, normally less than one per week (Utdanningsdirektoratet, 2006, 2007), and the English input to which students are exposed is thus necessarily limited. Furthermore, the Norwegian early English classroom typically does not provide very much L2 input, since it largely uses Norwegian as the language of instruction. One reason for this situation may be that English is not an obligatory subject in teacher training in Norway, although most teachers in lower primary school will normally have to teach the subject (cf. Guldal in Trønder-Avisa, 2007). Also, the curriculum and teaching materials commonly used in the classroom reflect a teaching style where the target language is the object, but not the medium of instruction.

Vocabulary acquisition in an L2 has traditionally been associated with rote learning and memorization of words (cf. Kersten, 2010). However, L2 vocabulary can obviously also be acquired from naturalistic input, as is the case in L1. In fact, vocabulary acquisition may not be subject to age effects. While it is the area of language first evident in young children, we all continue to learn new words throughout life. We know that vocabulary is acquired at a fast pace in school (see e.g., Nagy and Herman, 1987; Clark, 1993; Pinker, 1994; Bloom, 2000, 2004; Berman, 2007). On the other hand, vocabulary is an aspect of language for which L1 and L2 acquisition may be assumed to differ. In L1, vocabulary acquisition entails the daunting task of learning concepts and words at once. The L2 learner, on the other hand, will generally have acquired the concepts already. Many theories have been proposed about bilingual vocabulary acquisition, some involving the L1 as a mediator, while others assume a direct link to the concept. Without engaging in a discussion of the extent to which cross-linguistic lexical variation reflects deeper conceptual differences, we assume that L2 vocabulary acquisition, at least at early stages, and at least when the L1 and the L2 represent similar cultures, does to a large extent entail learning the new labels for familiar concepts (see e.g., Singleton, 1999, p. 48; MacWhinney, 2005).

There is thus no reason to believe that neither age nor the presence of an already acquired L1 should have a detrimental effect on vocabulary acquisition, and we should expect that increased exposure to an L2 during the first year of school will lead to naturalistic acquisition and significantly increased vocabulary comprehension. Specifically, it is likely that input alone is particularly beneficial for vocabulary acquisition in young L2 learners. Shintani (2011) explicitly investigated whether input-only instruction may be as effective as production-based instruction for 6–8-year old Japanese learners of English, hypothesizing that mechanisms such as fast mapping are still available at this age. The conclusions of her study are indeed that in this age group, the effect of input-based instruction on vocabulary acquisition is as good as, or better than, that of production-based instruction.

METHOD

The present study investigates whether employing a bilingual approach to an otherwise normal Norwegian first-grade English classroom will lead to improved acquisition over 1 year, compared to a standard, i.e., largely native language-based, first grade class. The research questions are whether children in each of the classes improve significantly in vocabulary acquisition over their first year of school, and whether there is a measurable difference in the two groups' vocabulary comprehension at the end of the first grade.

CONDITIONS

Two different schools were recruited for the experiment. In one school, teachers were told to do nothing out of the ordinary, and to teach English to their first-graders the way they would normally do, with the L1 as the main medium of instruction. In the other school, teachers agreed to use English more extensively with the children in and outside of English class, such as for morning meetings, simple instructions during the day, and reading aloud. However, they were not instructed to avoid the use of the L1; this school's approach to English teaching can thus be said to be bilingually-based.

The two schools were both standard state schools, situated in similar suburban areas in one of Norway's largest towns. The areas from which the schools recruit their pupils are socioeconomically comparable; they are both relatively affluent, with mean incomes slightly above the national average. The ethnic makeup of the two neighborhoods is also comparable, with a low percentage of families with immigrant backgrounds. On the national tests of English for 5th grade in 2008 the two schools scored similarly at or (in the case of the native language-based classroom's school) a little above the average. Thus, there is every reason to believe that these two schools are comparable in terms of student population and quality of teaching, and that they are representative of Norwegian

state schools. In addition, a parent questionnaire asked for background information about the children concerning factors such as foreign travel, English-speaking friends and relatives, and input received through media. None of the participants included in the study had extended stays beyond normal vacations abroad, and none had close English-speaking family or other special circumstances which might make their English competence atypical for a Norwegian 6-year-old. Although the parental reports were relatively crude, information was quantified by counting weeks spent outside of Scandinavia and hours per week with English exposure from games and media prior to starting school. This information is summarized for both groups in **Table 1**.

A Mann-Whitney *U* test found no significant difference for weeks spent outside Scandinavia [$U_{(59)} = 399.0$, $Z = -0.751$, $p = 0.453$], nor on previous exposure to English [$U_{(59)} = 407.0$, $Z = -0.652$, $p = 0.515$]. It thus seems safe to assume that these children's English exposure outside of school was similar.

Children in the two groups were also similar on a number of factors that may potentially influence English acquisition, which will be more closely described in the test materials section.

In each school, three different classes and class teachers participated in the project. In the bilingually-based school, one teacher had the main responsibility for English classes in all groups. In this school, groups were often organized across classes for various subjects, and this teacher was a natural choice for English classes since she was a native speaker of English. However, all class teachers participated in providing input throughout the school day. In each school, one teacher was responsible for recording information on time spent on English, and about activities and materials used, and to report to the researcher. These reports were frequent and relatively informal during the two periods of test sessions in September and May. In the middle of the spring term, both teachers formally reported on the same three questions (time, activities, and materials) in emails to the researcher. Information from both schools indicated that they consistently followed the pattern described below throughout the test period.

The native language-based condition school reported formally spending 30 min a week on English class. They also reported spending a few minutes in morning meetings every day talking about the weather and the names of the days in English, but these meetings were otherwise conducted entirely in Norwegian. The time spent on English in this group was thus of around 45 min per week, which is representative of the average that normal Norwegian schools spend in the first grade. Also representative is the fact that communication during this time took place mainly in Norwegian. Activities in this group included the use of the workbook Junior Scoop 1–2 (Bruskeland and Ranke, 2005) which is intended for use in first grade, and which contains simple activities, including routine instructions, rhymes and songs. Furthermore, teachers reported a number of other English songs used in class.

The bilingually-based group spent about 30–40 min per week on English class. While this school also uses the Scoop series of work- and textbooks, it was not used in this group of children. The teachers instead used various other materials, including simple stories and books, which the teacher read aloud, often with illustrations. Teachers would also spend time talking about pictures or objects. Furthermore, this group spent more time speaking English during morning meeting time; the teacher estimated about 5–10 min per day. While the native language-based group's morning meetings were conducted in Norwegian, with only routine discussion of words for the weather and the days of the week in English, morning meetings in the bilingually-based group were more or less conducted in English on the part of the teacher, while the pupils were free to answer in either language as they wished. Teachers in the bilingually-based school also used English for simple classroom management throughout the day, often with Norwegian translations instantly following, such as the reminder "No running in the corridors—*ikke løpe i gangen*!."

It is important to point out, then, that the change in the English classroom of the bilingually-based group did not consist of more formal instruction or an increase in teaching hours for English. Time spent on English was a little higher than is normal in Norwegian schools, but with an average of around 70 min per week including morning meetings, it still is a small proportion of the total time spent at school. Furthermore, there was no focus on pupils' production, even though increased L2 production may have been a natural consequence of the increased input. In other words, the change in this school consisted solely of an increased focus on providing target language exposure in a natural context.

Table 1 | Mean, minimum and maximum values and standard deviations (*SD*) for weeks spent outside of Scandinavia and hours of exposure from media, games and music prior to starting school in the bilingually-based and the native language-based groups.

	Weeks outside Scandinavia		Exposure from media, games, music	
	BB	**NB**	**BB**	**NB**
Min	0.0	0.0	0.0	0.0
Max	14.0	24.0	11.0	7.0
Mean	4.2	6.4	1.8	1.2
SD	3.6	6.8	2.5	1.6

BB, bilingually-based; NB, native language-based.

PARTICIPANTS

All parents of students in the relevant first grades were contacted in writing and asked for written consent for their child to participate. Approximately 80% of the parents provided consent in each group. In the bilingually-based group, the total number of volunteers was 59. 10 participants were excluded because they were bilingual, and one because he had participated in another, related study. From the remaining 49 children, 31 were randomly selected for the project by the researcher. The final test group consisted of 17 boys and 14 girls, all monolingual speakers of Norwegian with no known diagnosis which might influence acquisition. In the native-language based group there were 35 volunteers. Three were excluded because of bilingualism, and one because of hearing problems. Two children participated in the pre-test only; one because he was not available during the post-test, the other because he did not want to participate in it. The final test group

consisted of 15 boys and 14 girls, all monolingual and with no known diagnosis which may have had consequences for the study.

Mean age at the time of pre-testing was 6;1 in both groups, with no significant difference [$U_{(59)} = 433.0$, $Z = -0.245$, $p = 0.806$]. The project was approved by the Norwegian Social Science Data Services (NSD). **Table 2** summarizes descriptive statistics for age and scores on background measures in the two groups.

TEST MATERIALS

English vocabulary comprehension was tested using Form B of the Peabody Picture Vocabulary Test, Fourth Edition (PPVT™-4). This test measures vocabulary comprehension by means of pictures; the subject hears a word and selects the corresponding picture from a set of four options. This means that issues related to literacy can be avoided, and no L2 production is necessary on the part of the participant. Both these criteria made the test particularly well suited to these young learners, whose level both of literacy and of English was too low, especially in the pre-test, for more comprehensive tests to yield reliable results.

Pre-testing took place within the first 6 weeks of the children's 1st school year. During the pre-test session Norwegian vocabulary comprehension was tested in addition to initial English vocabulary comprehension, using a translated version of Form A of the Peabody Picture Vocabulary Test, Fourth Edition (PPVT™-4). There were no significant differences between the groups on either of these tests.

Post-testing took place during the last 6 weeks of the school year. During this test session, in addition to the post-test of English vocabulary comprehension, visio-spatial working memory was tested using a memory game where the child memorized sets of picture cards which were then turned face down, and was asked to find the pairs in as few attempts as possible. Furthermore, non-verbal intelligence was tested using the Matrices section of the Kaufmann Brief Intelligence Test, Second Edition (KBIT-2), and verbal intelligence using a version of the Riddles section of the KBIT-2 translated into Norwegian. These particular control measures, including L1 vocabulary, were chosen firstly to control for group differences on the outset, and secondly to provide measures that are believed to correlate with L2 acquisition. There are consistent findings in research suggesting that L2 language competence correlates highly with working memory, non-verbal intelligence, and, most importantly, L1 competence and skills (Colledge et al., 2002; Gathercole, 2006; e.g., Sparks et al., 2009; Dale et al., 2010; Hayiou-Thomas et al., 2012; Foyn et al., under revision). No significant differences between the two groups were found on any of the control measures; see **Table 3**.

Each test session was conducted at the child's school, during school hours or in the after-school program. Testing took place in a quiet room, with only the child, the researcher, and sometimes an assistant present. Each test session lasted for approximately 1 h. Most children were able to complete test sessions without signs of fatigue; if they did show signs of losing concentration, they were given a short break. Average time between pre- and post-testing was eight months in both groups.

RESULTS

Because the sample is relatively small (native language-based: $n = 29$, bilingually-based: $n = 31$) and because the data are not normally distributed, data were analyzed with non-parametric tests.

Results from the pre-test reveal that the children in general knew very little English when starting school. The mean raw score of the native language-based group was 23.72, which according to

Table 2 | Mean, minimum and maximum values and standard deviations (*SD*) for age, vocabulary, verbal and non-verbal intelligence and memory scores (raw) in the bilingually-based and the native language-based groups.

	Age at pretest		Norwegian vocabulary		English vocabulary (pre-test)		English sentences (pre-test)		Non-verbal intelligence		Verbal intelligence		Memory	
	BB	NB	BB	NB	BB	NB	BB	NB	BB	NB	BB	NB	BB	NB
Min	5;6	5;6	97	85	2	3	3	2	12	11	11.0	10.0	40	40.0
Max	6;6	6;5	157	145	56	61	8	10	32	38	24.0	26.0	59	62.0
Mean	6;1	6;1	119.9	113.8	25.4	23.7	5.7	5.2	18.7	17.7	16.6	16.1	46.3	48.0
SD	0.027	0.028	14.8	14.6	11.4	13.6	1.5	1.9	4.6	5.2	3.9	4.7	4.6	5.9

BB, bilingually-based; NB, native language-based.

Table 3 | Mann-Whitney *U*, *Z*, and *p* for between-groups comparison of vocabulary, verbal and non-verbal intelligence and memory in the bilingually-based and the native language-based groups.

	Norwegian vocabulary	English vocabulary (pre-test)	English sentences (pre-test)	Non-verbal intelligence	Verbal intelligence	Memory
Mann-Whitney U	352.000	363.000	344.000	333.000	389.500	375.500
Z	−1.443	−1.281	−1.402	−1.747	−0.891	−1.098
p (two-tailed)	0.149	0.200	0.161	0.081	0.373	0.272

the PPVT™-4 manual has an age equivalent for native English speakers of 2;4. The mean raw score in the bilingually-based group was 25.39, with a native age equivalent of 2;5. In short, these Norwegian children demonstrated English comprehension comparable to very young English-speaking children. Both groups' age equivalents are in fact below the chronological age for which the PPVT™-4 is normed, which has a lower bound of 2;6, although they are above the lower bound for age equivalents, which is 2;0. Competence was very similar between the two groups, even though the bilingually-based group did score slightly higher. An independent samples Mann-Whitney *U* test reveals that this difference is not significant [$U_{(59)} = 363$, $Z = -1.281$, $p = 0.20$]. This is confirmed by the PPVT™-4 Manual, which allows raw scores to be converted into growth scale value (GSV) scores. According to the manual (Dunn and Dunn, 2007b, p. 205), between chronological ages 2;6 and 12;0, a GSV point difference of eight is considered significant. The difference between an average score of 23.72 (GSV 84) and one of 25.39 (GSV 85) is only one point, and is thus not significant. **Figure 1** illustrates the scores on the pre-test and the post-test in the two groups.

After 8 months, the mean raw score on the PPVT™-4 had increased for both groups; to 29.14 for the native language-based group, and to 44.10 for the bilingually-based group. A Mann-Whitney *U* test found the difference between the two groups at this time to be significant [$U_{(59)} = 207.5, Z = -3.582, p < 0.01$], and this finding is confirmed by the PPVT™-4 Manual, since the difference between a mean of 29.14 (GSV 89) and a mean of 44.10 (GSV 101) is 12 points.

GROUP DEVELOPMENT

For the repeated-measures test, the Wilcoxon signed ranks test was used. For the native language-based group, the test did not reveal a significant difference between the pre- and post-tests, with $W_{(28)} = 143.50, Z = -1.356, p = 0.0875$ (one-tailed). This finding is confirmed by GSV scores; the difference between mean pre- and post-test GSV scores is only five points, from 84 to 89 GSV points.

For the bilingually-based group, the median score was significantly different from the pre-test to the post-test with $W_{(30)} = 19.50$, $Z = -4.479$, $p < 0.01$ (one-tailed). Again, the GSV scores confirm the finding, since the difference between mean pre-test (GSV 85) and post-test (GSV 101) score is 16 points, which, according to the PPVT™-4 Manual, is a significant difference (Dunn and Dunn, 2007b, p. 205). The effect size for the bilingually-based group was 0.8, indicating that the change in these pupils' average receptive vocabulary from the beginning to the end of the school year was not only significant but also substantial.

Successful L2 acquisition does not necessarily equal near-nativeness, but comparison to L1 acquisition may nevertheless be useful for purposes of illustration. One measure of the meaningfulness of the development in the bilingually-based group is illustrated in **Table 4**, which summarizes the results and their age equivalents, as given by the PPVT™-4 manual. Thus, the native language-based group's (non-significant) mean increase in receptive vocabulary translates into an equivalent of only 3 months' development in native English children, from age 2;4 to age 2;7.

The mean age equivalent of the bilingually-based group, however, has increased by 10 months in the course of an average time span of 8 months. This means that these L2 learners have, on average, been acquiring new words at a slightly faster rate than the average for children at the same stage of language development, who are acquiring English as their L1. The main difference is that, while this development on average takes place between ages 2;5 and 3;3 in English-speaking children, it took place between mean ages 6;1 and 6;9 in these L2 learners. This is quite an astonishing development, considering that the input to which these children have been exposed is still very limited compared to that of children acquiring their native language. The results thus clearly indicate that there is no inherent problem in the early-start foreign language classroom *per se* preventing it from being successful, at least not with respect to vocabulary development.

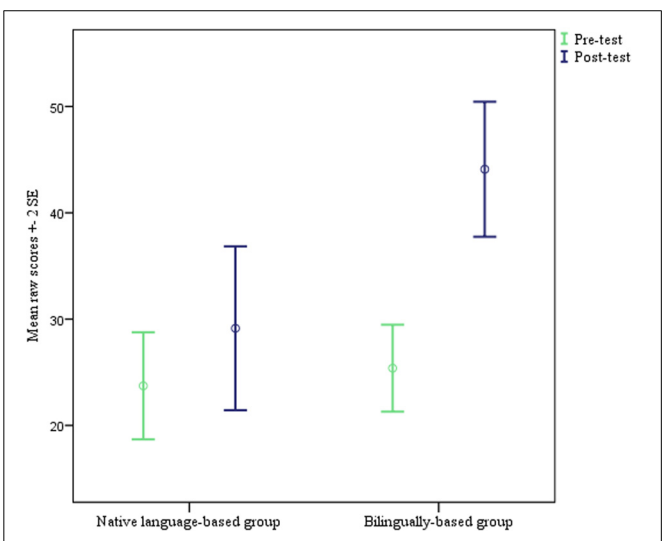


FIGURE 1 | PPVT-IV receptive vocabulary development, pre-test to post-test.

THE NATURE OF GROUP VOCABULARY DIFFERENCES

It is worth looking at group differences for cognates and non-cognates separately, since there may be differences in how the two categories are acquired. Because of the PPVT™-4 discontinuation rule, where for each set of 12 words, testing stops if the participant

Table 4 | Age equivalents of pre- and post-test vocabulary scores (raw) in the bilingually-based and the native language-based groups.

	PPVT™-4, pre-test		PPVT™-4, post-test	
	Mean score	L1 Age equivalent	Mean score	L1 Age equivalent
NB	23.72	2;4	29.14	2;7
BB	25.39	2;5	44.10	3;3

BB, bilingually-based; NB, native language-based.

makes eight or more errors, a few children were tested only on the 12 first words of the test in each session. These 12 words which all participants encountered are listed in **Table 5** below, followed by the percentages of children in each group who answered correctly for each word in the post-test, as well as the results of a Mann-Whitney *U* test comparing the two groups' responses for each word. Words that sound similar in the two languages (cognates) are given in bold.

The words which are successfully identified by virtually all children in both groups are *cat*, *apple*, *balloon*, and *hand*, all of which are phonologically similar to their Norwegian counterparts *katt, eple, ballong,* and *hand*. However, the bilingually-based group scores slightly higher also on these words; for *apple and hand*, the difference is significant. Furthermore, the words *tree* and *drinking* are recognized by virtually all the children in the bilingually-based group, and the difference between the two groups here is significant. These words also sound relatively similar to their Norwegian counterparts *tre* and *drikke*.

However, the children in the bilingually-based group outperform their native language-based group peers also on non-cognates. The percentage of children who correctly identify the words *airplane* and *bird*, whose Norwegian equivalents are *fly* and *fugl* respectively, is slightly higher in the bilingually-based group than in the native language-based group, although the difference is not significant, while the differences for the words *money* (Norwegian *penger*) and *umbrella* (Norwegian *paraply*) are significant, and the word *table* (Norwegian *bord*) is the one with by far the biggest difference in scores. While only 10% of the children in the native language-based group correctly identify this word, it is successfully identified by 90% of the children in the bilingually-based group. Although the number of items is too low to draw firm conclusions, we have an indication that the advantage in the bilingually-based group holds both for cognates and non-cognates.

Table 5 | Percentages of correct answers and Mann-Whitney U, Z, and p for between-groups comparisons of number of correct answers for cognate and non-cognate words in the bilingually-based and the native language-based groups.

Word	BB(%)	NB(%)	Mann-Whitney U (*df* = 59)	*Z*	Sig. (1-tailed)
Cat	100	100	341	0.000	0.500
Apple	100	93.1	310	−1.695	0.045
Balloon	100	89.7	321	−0.902	0.184
Hand	100	93.1	310	−1.695	0.045
Airplane	32.3	24.1	324	−0.386	0.350
Bird	32.3	27.6	335	−0.139	0.445
Tree	96.8	44.8	183.5	−3.516	0.000
Table	90.3	10.3	134.5	−4.304	0.000
Drinking	96.8	62.1	208	−3.311	0.001
Frog	61.3	55.2	313.5	−0.576	0.283
Money	67.7	44.8	249.5	−1.924	0.027
Umbrella	29	6.9	273	−1.747	0.041

BB, bilingually-based; NB, native language-based.

DISCUSSION

We see from the above results that English teaching in the native language-based group has had no significant impact on English receptive vocabulary. In other words, the 20+ h out of the 138 h of compulsory English teaching for grades 1–4 which this school is spending in the first grade have not had any measurable effect. We interpret this to mean that the L2 input received through this method of English teaching does not reach the critical threshold needed by children at this age for vocabulary development to take place. Children in both groups had acquired some English vocabulary prior to starting school, possibly through various sources such as computer games, music, TV, and movies. However, this vocabulary was very small for both groups, and included a number of cognates with Norwegian; word learning may have been incidental. The native language-based group's lack of English vocabulary development in the course of 8 months indicates that English exposure outside of school for young children in Norway is not sufficient for systematic acquisition. This further supports Murphy's (2010) argument that spending more time on the L2 in the classroom is especially important for learners who do not have extensive exposure to the target language outside of school.

We have also seen that the advantage in the bilingually-based group holds both for cognates and for non-cognates. Cognate and non-cognate acquisition may be slightly different processes in SLA. For example, Tonzar et al. (2009) show that cognates are acquired more easily than non-cognates both for English and German in Italian learners. Gascoigne (2001) proposes that cognates are in fact retrieved differently from other words in the L2, and that the representations in the two languages in the mental lexicon are partly overlapping. Aukrust (2007) argues that observed differences in whether vocabulary size in the two languages is correlated in bilingual children may be the result of how closely related the languages in question are, and consequently how many cognates there are that can be more or less transferred from one language to the other. Norwegian and English are both Germanic languages, and thus relatively closely related. Although a great portion of the English vocabulary is of Romance origin, basic words are more often Germanic, and a possible hypothesis could have been that the improved performance of the bilingually-based group is mainly a result of cognate comprehension. However, we saw in **Table 5** that the children in the bilingually-based group seem to outperform the native language-based group both on words which are cognates in Norwegian and English and on words which are not.

Another question is what the exact problem is for acquisition in the native language-based group. There are two alternative explanations for their lack of measurable development on the PPVT™-4 in the course of 8 months. The first possible explanation is that the vocabulary items tested were not frequent enough in this group's input, while the second is that the words tested in the PPVT™-4 were not present in the input at all. The difference is in whether the input in the native language-based group is best described as generally impoverished, or whether it is just naturally more specialized due to being more limited. The early words in the PPVT™-4 are those expected to be familiar to very young American children, and it is not obvious that these are the same as those emphasized in early Norwegian English classrooms. Since

the more limited input of the native language-based group necessarily includes a smaller number of words, it is possible that the children in this group have not happened to come across many of the words of the PPVT™-4, even if they may have acquired other words. This could give them an unfair disadvantage in the test. However, a look at the teaching materials and the activities reported for the native language-based group indicates that the vocabulary in the PPVT™-4 has been used in the classroom. Out of the (very few) words to be found in the native language-based group's work book *Junior Scoop 1–2*, several can be found in the early sets of the PPVT™-4, such as *tree*, *bird*, and *balloon*. Another area of vocabulary which teachers in this group specifically mentioned practicing was body parts, an area also present early on in the PPVT™-4 in words such as *hand* and *neck*. It is obviously impossible to establish whether the children have encountered all the words tested early in the PPVT™-4. Still, there is no reason to believe that the words in the PPVT™-4 are thematically different from those used in the native language-based classroom, and that this has created an unfair advantage for the bilingually-based group in the test.

Since the aim of the present study was to investigate the effect of an increase in English input which actually felt manageable and natural to the teachers in the bilingually-based classroom, the researcher did not interfere with what happened in the classroom, and the structure and nature of input was therefore not carefully controlled. This lack of control potentially raises questions about whether it is the exposure *per se* or specific characteristics of it which have brought about the effect. For example, one of the teachers in the bilingually-based school was in fact a native speaker of English, although she was also completely fluent in Norwegian and taught all other subjects in this language. It is of course conceivable that this teacher's nativeness in itself is what led to increased acquisition in the bilingually-based group. However, this would mean assuming that native speakers are always better teachers of L2s or that language can only successfully be acquired from native-speaker input, which goes against research findings both on L1 development (e.g., Singleton and Newport, 2004) and on second/foreign language teaching (cf. Moussu and Llurda, 2008). The main benefits of the teacher's nativeness, i.e., language proficiency and the confidence to use English extensively, can both be trained also in non-native teachers. It is precisely the conclusion of this paper that teachers should be trained in this.

Another question is whether the native language-based group really is representative of normal Norwegian schools, or whether the lack of acquisition in this group is a result of "poor" teaching. However, as already mentioned, pupils from this school have been previously found to perform above average in national tests in English. Whereas results in these tests may have come from classes taught by teachers other than those in the present study, it is highly unlikely that the school standards of English instruction have dramatically dropped. Furthermore, as with the bilingually-based group, the native language-based teachers knew that their pupils would be tested after 8 months, and were naturally eager for them to do well. If anything, it is likely that they spent more time on English than they would have in a normal year. Finally, and significantly, there is nothing in what the native language-based teachers report that deviates from the stated norms of the curriculum. As with many early-start foreign language programs, nothing in the plans for early English teaching in Norway focuses on extensive input for vocabulary acquisition.

CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH

The overall conclusion of the present study is that there is nothing inherent in the classroom situation which prevents successful L2 acquisition in young learners, and that vocabulary can be acquired at a fast rate in an early-start foreign language program. Furthermore, the study indicates that although such acquisition critically depends on input, exposure to the target language need not be unrealistically massive for acquisition to take place.

The PPVT™-4 only investigates receptive vocabulary, and tells us nothing about the productive vocabulary of the children in the study. However, we do know that the two are related, and that receptive vocabulary is important for comprehension, which in turn means that a larger receptive vocabulary allows more advanced input to be processed and understood. In this sense, receptive vocabulary can be assumed to be a predictor for further language acquisition.

A natural next step is to further examine whether such an increase in exposure to the target language has a long-term effect beyond the first year of school, and whether it is also evident in areas other than vocabulary comprehension. Furthermore, more research is needed concerning exactly what kind of input is necessary, including what proficiency level teachers must have attained and whether native input from sources other than the teacher, especially media (i.e., audio and video) can fruitfully be exploited to increase input in early-start second language classrooms.

AUTHOR CONTRIBUTIONS

Anne Dahl has had main responsibility for the project, including research design, data collection, analysis and interpretation, and drafting and revising the paper. Mila D. Vulchanova has contributed substantially to the conception and design of the research, and to critical revision of the paper for important intellectual content. Both authors have final approvement of the version to be published and agree to be accountable for all aspects of the work.

ACKNOWLEDGMENTS

We acknowledge the support of the Faculty of Humanities, Norwegian University of Science and Technology, from whom the first author received a grant to conduct this research.

REFERENCES

Abello-Contesse, C., Chacón-Beltrán, R., López-Jiménez, M. D., and Torreblancaa-López, M. M. (2006). "Introduction and overview," in *Age in L2 Acquisition and Teaching*, eds C. Abello-Contesse, R. Chacón-Beltrán, M. D. López-Jiménez, and M. M. Torreblancaa-López (Bern: Peter Lang), 7–27.

Aukrust, V. G. (2007). Young children acquiring second language vocabulary in preschool group-time: does amount, diversity, and discourse complexity of teacher talk matter? *J. Res. Child. Educ.* 22, 20. doi: 10.1080/02568540709594610

Berman, R. A. (2007). "Developing linguistic knowledge and language use across adolescence," in *Blackwell Handbook of Language Development,* eds E. Hoff and M. Shatz (Malden, MA: Blackwell), 347–367.

Bley-Vroman, R. (1989). "What is the logical problem of foreign language learning?," in *Linguistic Perspectives on Second Language Acquisition,* eds S. Gass and J. Schachter (Cambridge, MA: Cambridge University Press), 41–68.

Bloom, P. (2000). *How Children Learn the Meanings of Words.* Cambridge, MA: MIT Press.

Bloom, P. (2004). "Myths of word learning," in *Weaving a Lexicon,* eds D. G. Hall and S. R. Waxman (Cambridge, MA: MIT Press), 205–224.

Bruskeland, P. A., and Ranke, C. T. (2005). *Junior Scoop 1-2.* Oslo: Samlaget.

Burstall, C. (1975). French in the primary school: the British experiment. *Can. Mod. Lang. Rev.* 31, 388–402.

Cenoz, J. (2003). "The influence of age on the acquisition of English: general proficiency, attitudes and code-mixing," in *Age and the Acquisition of English as a Foreign Language,* eds M. García Mayo and M. García Lecumberri (Clevedon: Multilingual Matters), 77–93.

Childers, J. B., and Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Dev. Psychol.* 36, 11. doi: 10.1037/0012-1649.38.6.967

Clark, E. V. (1993). *The Lexicon in Acquisition.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511554377

Colledge, E., Bishop, D. V., Koeppen-Schomerus, G., Price, T. S., Happe, F. G., Eley, T. C., et al. (2002). The structure of language abilities at 4 years: a twin study. *Dev. Psychol.* 38, 749–757. doi: 10.1037//0012-1649.38.5.749

Dale, P. S., Harlaar, N., Haworth, C. M., and Plomin, R. (2010). Two by two: a twin study of second-language acquisition. *Psychol. Sci.* 21, 635–640. doi: 10.1177/0956797610368060

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Stud. Second Lang. Acquis.* 22, 499–533.

DeKeyser, R. M., and Larson-Hall, J. (2005). "What does the critical period really mean?," in *Handbook of Bilingualism: Psycholinguistic Approaches,* eds J. F. Kroll and A. M. B. De Groot (Oxford: Oxford University Press).

Dunn, L. M., and Dunn, D. M. (2007a). *Peabody Picture Vocabulary Test*. 4th Edn. Minneapolis, MN: Pearson Education, Inc.

Dunn, L. M., and Dunn, D. M. (2007b). *Peabody Picture Vocabulary Test Manual.* Minneapolis, MN: Pearson Education, Inc.

Felix, S. W. S. (1985). More evidence on competing cognitive systems. *Second Lang. Res.* 1, 47–72. doi: 10.1177/026765838500100104

García Lecumberri, M. L., and Gallardo, F. (2003). "English FL sounds in school learners of different ages," in *Age and the Acquisition of English as a Foreign Language,* eds M. D. P. García Mayo and M. L. García Lecumberri. (Clevedon: Multilingual Matters), 115–135.

García Mayo, M. D. P. (2003). "Age, length of exposure and grammaticality judgements in the acquisition of English as a foreign language," in *Age and the Acquisition of English as a Foreign Language,* eds M. L. García Lecumberri and M. D. P. García Mayo (Clevedon: Multilingual Matters), 94–114.

Gascoigne, C. (2001). Lexical and conceptual representations in more- and less-skilled bilinguals: the role of cognates. *Foreign Lang. Ann.* 34, 7. doi: 10.1111/j.1944-9720.2001.tb02084.x

Gass, S. M. (2003). "Input and interaction," in *The Handbook of Second Language Acquisition,* eds C. J. Doughty and M. H. Long (Oxford: Blackwell), 224–255.

Gathercole, S. E. (2006). Nonword repetition and word learning: the nature of the relationship. *Appl. Psycholinguist.* 27, 513–613. doi: 10.1017/S0142716406060383

Goldschneider, J. M., and DeKeyser, R. M. (2001). Explaining the "Natural Order of L2 Morpheme Acquisition" in English: a meta-analysis of multiple determinants. *Lang. Learn.* 51, 1–50. doi: 10.1111/1467-9922.00147

Hart, B., and Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore: Brookes.

Hayiou-Thomas, M. E., Dale, P. S., and Plomin, R. (2012). The etiology of variation in language skills changes with development: a longitudinal twin study of language from 2 to 12 years. *Dev. Sci.* 15, 233–249. doi: 10.1111/j.1467-7687.2011.01119.x

Hoff, E., and Naigles, L. (2002). How children use input to acquire a lexicon. *Child Dev.* 73, 418–433. doi: 10.1111/1467-8624.00415

Holmstrand, L. S. E. (1982). English in the *Elementary School: Theoretical and Empirical Aspects of the Early Teaching of English as a Foreign Language*. Uppsala: Acta Universitatis Upsaliensis.

Hyltenstam, K. (1992). "Non-native features of near-native speakers: on the ultimate attainment of childhood L2 learners," in *Cognitive Processing in Bilinguals,* ed R. J. Harris (Amsterdam: Elsevier Science), 351–368.

Hyltenstam, K., and Abrahamsson, N. (2003). "Maturational constraints in SLA," in *The Handbook of Second Language Acquisition,* eds C. J. Doughty and M. H. Long (Oxford: Blackwell), 539–588.

Johnson, J. S., and Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cogn. Psychol.* 21, 60–99. doi: 10.1016/0010-0285(89)90003-0

Kersten, S. (2010). The *Mental Lexicon* and *Vocabulary Learning: Implications for the Foreign Language Classroom.* Tübingen: Narr.

Larsen-Freeman, D. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Q.* 9, 409–419. doi: 10.2307/3585625

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Lang. Res.* 24, 35–63. doi: 10.1177/0267658307082981

Lasagabaster, D., and Doiz, A. (2003). "Maturational constraints on foreign-language written production," in *Age and the Acquisition of English as a Foreign Language*, eds M. L. García Lecumberri and M. D. P. García Mayo (Clevedon: Multilingual Matters), 136–160.

Lightbown, P. M. (2000). Classroom SLA research and second language teaching. *Appl. Linguist.* 21, 431–462. doi: 10.1093/applin/21.4.431

Lightbown, P. M., Halter, R., White, J., and Horst, M. (2002). Comprehension-based learning: the limits of "Do It Yourself." *Can. Mod. Lang. Rev.* 58, 427–464. doi: 10.3138/cmlr.58.3.427

MacWhinney, B. (2005). "A unified model of language acquisition," in *Handbook of Bilingualism: Psycholinguistic Approaches,* ed A. M. B. De Groot (Cary, NC: Oxford University Press), 49–67.

Moussu, L., and Llurda, E. (2008). Non-native English-speaking English language teachers: history and research. *Lang. Teach.* 41, 315–348. doi: 10.1017/S0261444808005028

Muñoz, C. (2001). "Factores escolares e individuales en el aprendizaje formalde un idioma extranjero," in *Estudios de Lingüística. Anexo 1: Tendencias y Líneas de Investigación en Adquisición de Segundas Lenguas*, eds S. P. Cesteros and V. S. García (Alicante: Universidad de Alicante), 249–270.

Muñoz, C. (2006). "The BAF Project: research on the effects of age on foreign language acquisition," in *Age in L2 Acquisition and Teaching,* eds C. Abello-Contesse, R. Chacón-Beltrán, M. D. López-Jiménez, and M. M. Torreblancaa-López (Bern: Peter Lang), 81–92.

Murphy, V. A. (2010). "The relationship between age of learning and type of linguistic exposure in children learning a second language," in *Continuum Companion to Second Language Acquisition,* ed E. Macaro. (London; New York: Continuum International Publishing Group), 158–178.

Nagy, W. E., and Herman, P. A. (1987). "Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction," in *The Nature of Vocabulary Acquisition,* eds M. G. Mckeown and M. E. Curtis (Hillsdale, NJ: Erlbaum), 19–36.

Newport, E. L. E. (1990). Maturational constraints on language learning. *Cogn. Sci.* 14, 11–28. doi: 10.1207/s15516709cog1401_2

Nikolov, M. (2009). "The age factor in context," in *The Age Factor and Early Language Learning,* ed M. Nikolov (Berlin: Mouton de Gruyter), 1–38.

Pelucchi, B., Hay, J. F., and Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Dev.* 80, 674–685. doi: 10.1111/j.1467-8624.2009.01290.x

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language.* New York, NY: Harper Perennial Modern Classics.

Ruiz-González, G. (2006). "Age effects on single phoneme perception," in *Age in L2 Acquisition and Teaching*, eds C. Abello-Contesse, R. Chacón-Beltrán, M. D. López-Jiménez and M. M. Torreblancaa-López (Bern: Peter Lang), 155–173.

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *J. Mem. Lang.* 35, 606–621. doi: 10.1006/jmla.1996.0032

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., and Barrueco, S. (1997). Incidental language learning: listening (and learning) out of the corner of your ear. *Psychol. Sci.* 8, 4. doi: 10.1111/j.1467-9280.1997.tb00690.x

Shintani, N. (2011). A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Lang. Teach. Res.* 15, 137–158. doi: 10.1177/1362168810388692

Singleton, D. (1999). *Exploring the Second Language Mental Lexicon.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524636

Singleton, D., and Ryan, L. (2004). *Language Acquisition: The Age Factor.* Clevedon: Multilingual Matters.

Singleton, J. L., and Newport, E. L. (2004). When learners surpass their models: the acquisition of American Sign Language from inconsistent input. *Cogn. Psychol.* 49, 370–407. doi: 10.1016/j.cogpsych.2004.05.001

Sparks, R. L., Patton, J. O. N., Ganschow, L., and Humbach, N. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Appl. Psychol.* 30, 725–755. doi: 10.1017/S0142716409990099

Stern, H. H. (1983). *Fundamental Concepts of Language Teaching.* Oxford: Oxford University Press.

Tonzar, C., Lotto, L., and Job, R. (2009). L2 vocabulary acquisition in children: effects of learning method and cognate status. *Lang. Learn.* 59, 23. doi: 10.1111/j.1467-9922.2009.00519.x

Trønder-Avisa. (2007). *Bekymret for engelskfaget [Online]*. Available online at: http://www.t-a.no/nyheter/article176154.ece#.UoEQ2OKmb30 (Accessed November 11, 2013).

Utdanningsdirektoratet. (2006). *Knowledge Promotion: Curriculum in English*. Available online at: http://www.udir.no

Utdanningsdirektoratet. (2007). *Ansvaret for fordeling av skoledager utover året*. Available online at: http://www.udir.no

Vulchanova, M., Vulchanov, V., Sarzhanova, D., and Eshuis, H. (2012). The role of input in early bilingual lexical development. *Lingue e linguaggio* 17, 181–198. doi: 10.1418/38785

Wode, H. (1981). *Learning a Second Language: An Integrated View of Language Acquisition*. Tuebingen: Narr.

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: 19 December 2013; accepted: 30 March 2014; published online: 17 April 2014.

Citation: Dahl A and Vulchanova MD (2014) Naturalistic acquisition in an early language classroom. Front. Psychol. ***5****:329. doi: 10.3389/fpsyg.2014.00329*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 09 January 2015
doi: 10.3389/fpsyg.2014.01510

As naturalistic as it gets: subtitles in the English classroom in Norway

Mila Vulchanova*, Lisa M. G. Aurstad, Ingrid E. N. Kvitnes and Hendrik Eshuis

Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Edited by:
Vicky Chondrogianni, University of Edinburgh, UK

Reviewed by:
Jon Andoni Dunabeitia, Basque Center on Cognition, Brain and Language, Spain
Rosa Alonso Alonso, University of Vigo, Spain

****Correspondence:***
Mila Vulchanova, Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology (NTNU), Edvard Bulls veg 1, 7491 Trondheim, Norway
e-mail: mila.vulchanova@ntnu.no

This study aimed to investigate the effects of subtitles in the context of authentic material on second language comprehension and potentially, second language acquisition for Norwegian learners of English. Participants in the study were 49 17-year-old students and 65 16-year-old students, who were all native speakers of Norwegian learning English as an L2 in high school. Both age groups were divided into three Conditions, where one group watched an episode of the American animated cartoon Family Guy with Norwegian subtitles, one group with English subtitles, and one group watched the episode with no subtitles. On a comprehension questionnaire conducted immediately after watching the episode positive short-term effects of both native language (L1) and target language (L2) subtitles were found for both age groups. However, no differences in terms of the language of the subtitles were found in the older and more advanced group. Four weeks later the participants responded to a word definition task and a word recall task to investigate potential long-term effects of the subtitles. The only long-term effect was found in the word definition task and was modulated by age. We found, however, that native language subtitles impact negatively on performance on the comprehension task. The results from this study suggest that the mere presence of subtitles as an additional source of information enhances learners' comprehension of the plot and content in animated audio-visual material in their L2. The absence of differences in terms of the language of the subtitles in the more advanced group suggests that both intralanguage and interlanguage subtitles can aid target language comprehension in very advanced learners, most probably due to better consolidated vocabulary knowledge in that group. The two groups differed also on predictors of performance on the two lexical tasks. While in the less proficient younger group, vocabulary status best predicted performance on both tasks (vocabulary predicts vocabulary), for the very advanced older group, grammar was a stronger predictor, highlighting the importance of generic language competence and skills in L2 tasks for highly proficient L2 users. We also found an effect of written L2 skills on performance on both lexical tasks indicative of the role of orthography in vocabulary consolidation.

Keywords: second language acquisition, subtitles, authentic target language input, comprehension, vocabulary

INTRODUCTION

In the Norwegian context, second language learning typically takes place in schools. However, learners are often exposed to input from the target language outside the school setting through television, movies, newspapers and the internet (Dahl and Vulchanova, 2014). Audio-visual material is a frequently used resource for teaching and learning English as an L2 and it provides learners with natural spoken dialog in the target language. Audio-visual material can be presented to learners without any subtitles, with native language (L1) subtitles, or with target language (L2) subtitles. From the point of view of SLA in a naturalistic environment, the question is what type of subtitles, if any, impact on learners' comprehension. The current study aimed to address the role of subtitled audio-visual material in L2 comprehension in the Norwegian context.

Linguistic input in the target language is essential for second language acquisition (Verspoor et al., 2009; Ellis, 2013), and exposure to such input can be obtained through reading and listening. When assessing the extent to which learners can benefit from exposure to L2 input one needs to take into account the learners' outset in terms of level of proficiency. There seems to be a consensus that an incipient learner cannot benefit from the same amount of input to the same extent as a more proficient learner due to processing limitations (Chiquito, 1995; Verspoor et al., 2009). Gilmore (2007) nevertheless argues that authentic input should not be adapted to the proficiency level of the L2 learner when exposing SLA learners to input in the target language. In this way, the target language is presented in a naturalistic manner, thus offering a far richer sample of the target language than adjusted materials (Gilmore, 2007; Benavent and Peñamaría,

2011). Authentic material, thus, provides learners with natural samples of the target language, facilitating the SLA process by advancing linguistic competence in the target language (Verspoor and Winitz, 1997). Audio-visual material can be a good source of authentic input, as suggested by Gilmore (2007), and even though authentic input can be argued to be too challenging for L2 learners (Day, 2003; Flowerdew and Peacock, 2011), the use of animated cartoons can be argued to provide authentic input in the classroom. Cartoons often involve more clearly enunciated speech in standard accents in the target language (Sherman, 2003). The bright colors and exaggerated intonation and other features can also increase the motivation of L2 learners, thus creating a better environment for learning (Sherman, 2003; Bahrani and Soltani, 2011). Bahrani and Soltani (2011) further recommended the use of animated cartoons in classroom activities, as they provide variation for the brain in engaging both the left and right hemisphere, and prevent the students from being bored.

Audio-visual material can be a particularly good source for authentic input, when accompanied by subtitles (Neuman and Koskinen, 1992; Baltova, 1999; Bianchi and Ciabattoni, 2008). The combination of auditory input in the L2, nonverbal visual information, and verbal visual (orthographic) information can be argued to contribute to a better SLA learning environment than when only two or only one of the three information channels are available (Baltova, 1999). The textual information serves as an extra source of linguistic input either in the L1 or the L2. In a number of eye-tracking studies, d'Ydewalle and Van de Poel (1999) and Danan (2004) note that readers automatically read the subtitles, whenever they are available, indicating that the auditory and the verbal textual information are processed in parallel. These studies suggest that subtitles do not necessarily hinder the processing of the auditory information (d'Ydewalle and Gielen, 1992). Furthermore, the eye-tracking study by Bisson et al. (2014) suggests that participants spend time attending to both the subtitles and the visual images, thus making use of both channels. Whether subtitles in the L1 (also called interlingual) or in the L2 (intralingual subtitles) in an auditory context are more facilitative has been debated, with experiments showing controversial results depending on the aspect of language being tested and the age and level of proficiency of the participants. A number of researchers (Vanderplank, 1988; Markham, 1999; Bird and Williams, 2002; Danan, 2004; Mitterer and McQueen, 2009; Gunderson et al., 2011; Vandergift, 2011; and Bianchi and Ciabattoni, 2008) have found that L2 subtitles are more facilitatory. It can be argued that intralingual (target language) subtitles are useful, since they allow the learner to map phonology directly onto orthographic representations, and thus enhance speech segmentation making processing and comprehension of the auditory material much easier (Bird and Williams, 2002; Mitterer and McQueen, 2009). Other results, however, (cf. Guillroy, 1998; d'Ydewalle and Van de Poel, 1999; Bianchi and Ciabattoni, 2008; Zarei and Rashvand, 2011) suggest that L1 subtitles are more facilitatory. The results of Bianchi and Ciabattoni (2008) are particularly interesting for the current study, as they found that L1 subtitles were more facilitatory for the less proficient learners, whereas L2 subtitles were more facilitatory for the more advanced learners in their study. They argue that this difference might be due to L1 subtitles being automatically processed, while L2 subtitles may require more advanced knowledge of the L2, in order to have a positive effect (Guillroy, 1998; Bianchi and Ciabattoni, 2008). Guillroy (1998) also argues that L2 subtitles cannot compensate for difficult vocabulary and fast speech in audio-visual material. Therefore, the long term effects of target language subtitles, and subtitles in general, is still unclear (Vandergift, 2011).

HYPOTHESES

In this study, we hypothesized that subtitles would enhance participants' performance both on comprehension and on vocabulary learning (word definition and word identification), thus indicating facilitatory short-term and long term effects of subtitles in language learning. If the subtitles were indeed found to have an effect, we further hypothesized that the younger and less advanced group (16-year-old students) would benefit more from the Norwegian (L1) subtitles, while the older and more advanced group (17-year-old students) would benefit more from the English (L2) subtitles.

MATERIALS AND METHODS

PARTICIPANTS

Two age-groups were recruited for the study, a group of 16-year-old students ($N = 65$; $F = 34$, $M = 31$) and a group of 17-year-old students ($N = 49$; $F = 24$, $M = 25$). All participants were monolingual native speakers of Norwegian and attended the same school in a big city in Norway. The 17-year-old group was assumed to have a higher level of proficiency in English, as this was a more homogenous group of students who had all chosen to study English as one of their high school specialization subjects, whereas for the younger group English was still a compulsory school subject. The study was conducted in the participants' high school classrooms at times when they normally received instruction in English, making the environment for the research more natural.

The participants were divided into three Conditions (sub-groups) for each age group using their original school class classification. This grouping, based on original class classification led to the experiments being conducted during different times of the day based on the school schedule, and also caused the gender distribution to vary somewhat between the different sub-groups. All the students in all the six original classes were encouraged to participate in the study. Prior to the study, informed parental consent was elicited from the parents of all participants allowing the students to participate in the study, since the students were all under the age of majority. The study was also approved formally by the school and the teachers. Participants from both age groups were divided into three conditions, where one group watched the episode with Norwegian subtitles (Norwegian subtitles group), one with English subtitles (English subtitles group), and one with no subtitles (Control group).

The particular purpose of the study was not revealed to either students, teachers, or parents; they were only informed that the study investigated second language acquisition. Students not willing to participate or absent on one of the testing days, were not included in the study and the analyses. After all the testing had been completed, some further participants were excluded due to

missing background information from the background questionnaire, and were dropped from further analyses of the results. Also participants who reported severe hearing or visual problems or other language related problems that might have had an effect on the results were excluded. Participants with an L1 that was not Norwegian were also excluded. All participant information was treated anonymously and according to the rules prescribed by the Norwegian Data Protection Board which approved the study.

MATERIALS AND PROCEDURE

In order to establish a baseline, prior to the study, all participants were tested in English grammar and vocabulary competence. The grammar test was the Cambridge Essential Grammar in Use Level Test (available at http://www.cambridge.org/other_files/Flash_apps/inuse/EssGramTest/EssGramIndex.htm). In this test the participant fills in a blank choosing the correct option from among four alternatives. The test comprises 50 sentences and each item targets key areas of grammar (word forms, verbs, and verb forms, parts of speech, adverbials, word order). The maximum score is thus 50. The vocabulary test estimates participants' vocabulary size on the basis of performance on 10 word definitions (http://dynamo.dictionary.com/placement/level). This test has 4 levels, elementary school level, middle school level, high school level and college and beyond level. The degree of difficulty changes for each next level. The test is designed for native speakers of English. A typical score at level "College and beyond" can range between 45 and 55,000 words. Participants took the Word Dynamo test twice (in October 2012) responding at level "Middle school," and an average score was, thus, calculated for each participant. Both tests were conducted on-line on the internet, and the scores from these two tests were recorded by an experimenter. In addition, participants completed a background questionnaire (Appendix 4 in Supplementary Material) requesting information about their linguistic background, and focusing on extra-curricular activities where English as a second language might be involved, thus offering information on factors (variables) which impact on the process of second language acquisition, and as such, might potentially influence performance on the tasks. This questionnaire provided information on self-assessed L2 skills (speaking, reading, writing, and listening) on a scale from "basic" (1) to "fluent" (4); frequency of reading and writing English; frequency of watching English films and cartoons; choice of subtitle viewing when watching English movies (English/Norwegian/no); frequency of playing computer games in English and whether they had watched the *Family Guy* series (an episode of which was the experimental video). Frequency responses were on a scale from "rarely" (1) to "every day" (5). For the purposes of the analyses, only the numerical values of these responses, and their means, were used. The background questionnaire, the comprehension questionnaire, and the word recall task were all in paper format, and the participants responded using a pen/pencil.

When all participants had responded to the grammar and vocabulary tests and the background questionnaire, they watched an episode of the American animated cartoon Family Guy. The episode lasted approximately 20 min and was only watched once. The dialog was in standard American and British accents and was believed to be comprehensible to the participants, and the plot was presumed to be fairly easy to follow. We decided to use an animated cartoon episode, as cartoons often include more enunciated speech, and are believed to be more easily understood by the audience (Sherman, 2003). Moreover, a cartoon was expected to motivate the participants into paying more attention, thus creating a reliable context for the study (Sherman, 2003; Bahrani and Soltani, 2011).

In order to investigate the potential short term effects of subtitles on the comprehension of the episode, the participants responded to a comprehension questionnaire in a multiple choice format immediately after watching the episode (Appendix 1 in Supplementary Material). Four weeks later the participants responded to a word definition task in a multiple choice format (Appendix 2 in Supplementary Material) and a word recall task (Appendix 3 in Supplementary Material), both including words and phrases the participants had encountered in the Family Guy episode they had watched previously. This was done in order to investigate potential long term effects of the subtitles. The word definition task was administered in a multiple choice format and consisted of 30 words and phrases that occurred in the episode. In this task, participants were asked to select the correct definition from four alternatives. The word recall task consisted of a list of 53 words of which 22 occurred in the episode and 31 did not occur in the episode (which acted as distractors). The 22 target items selected for the word recall task were all semantically related to the plot of the episode. Thus, our expectation was that there was a likelihood that participants would recognize them, if they had processed the episode at a deeper (semantic) level. Participants were asked to identify the words they believed had occurred in the episode. The frequency of the words in both the word definition task and the word recall task was established in the Corpus of Contemporary American English (COCA) and was included as a potential factor in the inferential analysis. For all three tasks, instructions were written on top of the paper as well as given orally by the experimenters.

ANALYSES AND RESULTS

Inferential statistics was conducted in R using a generalized linear mixed model fit by the Laplace approximation to check for dependencies between the results from the tests, the presence/absence and nature of subtitled stimuli and variables from the participants' background (Baayen et al., 2008; Bolker et al., 2009). The models in R were created by using the factors from the initial background testing and the vocabulary and grammar test results. The results from the comprehension questionnaire, the word definition task, and the word recall task were analyzed independently (coded as correct or incorrect response for each item). Age and subtitle condition were included as between subject factors, test item and subject were included as random effects, while word frequency and the results of the grammar test, the vocabulary test and the background questionnaire were included as covariates. The models created in R were compared using likelihood ratio tests (ANOVA) in order to find the best fitting models. The best

fitting models for each of the tasks are presented in the Results Section below. The ratio between the different groups varied from test to test, and the models will therefore have different variables as predictors of the results.

RESULTS

Initial L2 proficiency results

The average scores on the grammar and vocabulary tests are presented in **Tables 1**, **2**. On the grammar test, the scores were calculated as the number of correct responses out of 50, and on the vocabulary test the scores were calculated as the estimated amount of English words known to the participants (see http://dynamo.dictionary.com/placement/level).

Tables 1, **2** show that the expected difference in L2 proficiency between the 16-year-old group and the 17-year-old group was justified. Overall, the 16-year-old students on average achieved lower scores on both the grammar and the vocabulary tests [ANOVA: $F_{(1,\ 108)} = 7.467,\ p < 0.01$ resp. $F_{(1,\ 108)} = 20.235,\ p < 0.001$]. For the vocabulary test also a significant interaction between Age and Subtitle Condition was found [$F_{(2,\ 108)} = 5.755, p < 0.005$]. Separate analyses for the Age groups revealed no significant differences for the 17-year-olds. For the 16-year-olds, participants in the Norwegian subtitles Condition significantly outperformed the Control and the English subtitles groups on the vocabulary test (Bonferroni corrected: $p < 0.01$ resp. $p < 0.05$). The high score of participants in this condition will be considered in more detail later.

Comprehension questionnaire

Positive short term effects of the subtitles were found for both the 16-year-olds and the 17-year-olds in the analysis of the comprehension questionnaire. The results from the comprehension questionnaire are shown in **Tables 3–5** below (for each analysis the best-fitting model is shown; the number of included factors accordingly varies).

As can be seen in **Table 3**, there is only marginally a difference between the 16 and 17-year-olds with the older group giving more correct responses (level of significance $p < 0.1$). There are clear short term effects of the subtitles ($p < 0.001$). The availability of subtitles for participants in both the English and the Norwegian subtitles Conditions enhanced these participants' performance, indicating positive short term effects of the subtitles on comprehension of the contents of the episode. Interestingly, however, the analysis indicates that the language of the subtitles did not matter, as the effect of subtitles in both conditions is highly significant. This suggests that both intralingual and interlingual subtitles as a source of input facilitated the comprehension of the episode, with these two groups performing significantly better than the control group. The most significant factor, however, was the computer game factor, estimated as the amount of time the participants had spent playing English computer games ($p < 0.001$). Also, the results from the grammar test were significant ($p < 0.05$), suggesting that higher grammar competence, as indicated by better performance on the grammar test, was a reliable predictor of comprehension of the Family Guy episode. Likewise, the score on the vocabulary test ($p < 0.01$) significantly predicted performance on the comprehension task. Interestingly, having—in daily life—the habit of displaying Norwegian subtitles when watching English films and series is associated with lower scores on the task ($p < 0.05$), while displaying English subtitles increases performance (n.s.) as compared to not displaying subtitles at all. On the other hand, the amount of time that participants usually spend on watching animated cartoons in English ($p < 0.05$) decreases scores as does the more frequent watching of English films and series ($p < 0.05$). The composite factor created

Table 1 | Average score on pre-study grammar (Cambridge Essential Grammar in Use Test) and vocabulary tests (Word Dynamo) (17-year-old students).

Condition (group)	*N*	Vocabulary	Grammar
Norwegian subtitles	14	14 747.79	45.50
English subtitles	16	15 499.31	46.13
Control group	19	16 917.53	45.21

Table 2 | Average score on pre-study grammar (Cambridge Essential Grammar Test) and vocabulary tests (Word Dynamo) (16-year-old students).

Condition (group)	*N*	Vocabulary	Grammar
Norwegian subtitles	16	14 842.91	44.88
English subtitles	25	11 658.72	42.68
Control group	24	11 015.65	44.67

Table 3 | Results of the inferential analysis of the comprehension questionnaire (both age groups).

	Estimate	*SE*	*z*-value	Pr(>|z|)
(Intercept)	−28.37734	5.71280	−4.967	6.79e-07***
Age	0.41712	0.24820	1.681	0.092854(*)
GroupEng	1.15136	0.26492	4.346	1.39e-05***
GroupNor	1.09320	0.29713	3.679	0.000234***
log(Voc)	1.19687	0.39036	3.066	0.002169**
log(Grammar)	2.95033	1.38667	2.128	0.033367*
FilmSubeng	0.60008	0.47325	1.268	0.204803
FilmSubnorw	−0.62278	0.24509	−2.541	0.011054*
Cartoon	−0.18873	0.08522	−2.215	0.026781*
Eng_game	0.40576	0.08350	4.859	1.18e-06***
Film_EngT	−0.32238	0.15565	−2.071	0.038346*
Read_Norw	0.29916	0.16017	1.868	0.061804(*)
Listen	0.91757	0.37791	2.428	0.015183*

*Age, Age group; GroupEng, English subtitles Condition; GroupNor, Norwegian subtitles Condition; log(voc), vocabulary test results (log); log(Grammar), grammar test results (log); FilmSubeng, English subtitles when watching an English movie; FilmSubnor, Norwegian subtitles when watching an English movie; Cartoon, amount of time spent watching animated cartoons in English; Eng_Game, amount of time spent playing English computer games; Film_EngT, frequency of watching English films and series; Read_Norw, self-estimated Norwegian reading skills; Listen, composite factor created by the ratio of self-estimated English listening skills compared to self-estimated English speaking skills. Significance codes: (*) p ≤ 0.1; * p ≤ 0.05; ** p ≤ 0.01; *** p ≤ 0.001.*

Table 4 | Results of the inferential analysis of the comprehension questionnaire (17-year-old group).

	Estimate	*SE*	*z*-value	Pr(>|z|)
(Intercept)	−35.6168	13.9389	−2.555	0.01061*
GroupEng	1.5088	0.4979	3.030	0.00244**
GroupNor	1.5172	0.5004	3.032	0.00243**
EngGame	2.3338	0.5305	4.399	1.09e-05***
log(Grammar)	9.3537	3.6623	2.554	0.01065*
FGyes	0.6939	0.4045	1.716	0.08623(*)

*GroupEng, English subtitles Condition; GroupNor, Norwegian subtitles Condition; EngGame, composite factor created by the ratio of the amount of time spent playing English computer games compared to self-estimated English writing skills; log(Grammar), grammar test results (log); FGyes, have watched Family Guy before. Significance codes: (*) $p \leq 0.1$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.*

Table 5 | Results from analysis of comprehension questionnaire (16-year-old group).

	Estimate	*SE*	*z*-value	Pr(>|z|)
(Intercept)	−12.74289	3.50561	3.635	0.000278***
GroupEng	0.78831	0.29695	2.655	0.007938**
GroupNor	0.75125	0.36342	2.067	0.038722*
log(voc)	1.33337	0.38792	3.437	0.000588***
listen_eng	0.65233	0.20861	3.127	0.001766**
Eng_game	0.18479	0.09895	1.868	0.061831(*)

*GroupEng, English subtitles Condition; GroupNor, Norwegian subtitles Condition; listen_eng, self-estimated English listening skills; log(voc), vocabulary test results (log); Eng_game, amount of time spent playing English computer games. Significance codes: (*) $p \leq 0.1$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.*

by the ratio of self-estimated English listening skills compared to self-estimated English speaking skills is tied to increased scores ($p < 0.05$). Finally, the self-estimated Norwegian reading skill is only marginally predicting results ($p < 0.1$).

The analysis of the results from the comprehension questionnaire for the 17-year-old group only indicates clear short term effects of the subtitles (level of significance $p < 0.01$). The availability of subtitles for participants in both the English and the Norwegian subtitles Conditions enhanced these participants' performance, indicating positive short term effects of the subtitles on comprehension of the contents of the episode. For that specific age-group, too, the analysis indicates that the language of the subtitles did not matter, as the effect of subtitles in both conditions is equally significant (almost identical *p*-values for the two Conditions/groups). The most significant factor in that age-group, however, was the composite computer game factor, estimated as the amount of time the participants had spent playing English computer games, compared to how proficient they estimated themselves to be at writing English from the L2 skill self-assessment part of the background questionnaire ($p < 0.001$). Also, the results from the grammar test were significant ($p < 0.05$), suggesting that higher grammar competence, as indicated by better performance on the grammar test, was a reliable predictor of comprehension of the Family Guy episode. The amount of time the participants had spent watching Family Guy before was also marginally significant ($p < 0.1$).

For the 16-year-old age group, participants in the English subtitles Condition performed significantly better than the two other groups ($p < 0.01$). Norwegian subtitles also seemed to enhance comprehension of the Family Guy episode ($p < 0.05$) for participants in that condition, however, this result should be seen in relation to this group's performance on the vocabulary test. The most significant predictor of comprehension in the 16-year-old group as a whole was the score on the vocabulary test ($p < 0.001$), something which might explain why the Norwegian subtitles were less significant than the English subtitles, since the Norwegian subtitles group performed better on the vocabulary task. Self-estimated English listening skills were also an important predictor ($p < 0.01$), suggesting that participants who were good at processing English aural material, were also good at understanding the video, and most likely used both the subtitles and the audio to process the contents. This factor also emerged as significant in the combined group analysis. In addition, the amount of time spent playing English computer games marginally predicted the results ($p < 0.1$).

Word definition task

We ran a combined analysis including Age (age group: 16|17) as an interacting factor. The results of the analysis are presented in **Table 6** below. A fixed main effect was found for Grammar, English writing skills, and video viewing Condition, with native language subtitles facilitating performance on the task, but this effect was modulated by age. In addition, there was a negative effect of the habit of watching films with native language subtitles (as reported in the questionnaire). No effects of the subtitles were, however, found on the word definition task in separate age group analyses, indicating that there were no long term effects for each group independently of the type of subtitles included in the different experimental conditions. Other factors were, however, found to influence the participants' performance. The results from the inferential analysis are shown in **Tables 7**, **8**.

The presence of subtitles did not predict performance of the 17-year-old participants on the word definition task. The most significant factor was, as in the comprehension questionnaire, the (self-reported) amount of time spent playing English computer games ($p < 0.001$). Also, the grammar test results predicted the participants' performance ($p < 0.01$). The self-estimated speaking skills of the participants were also marginally significant ($p < 0.1$). Finally, quite surprisingly, the analysis marginally indicated that the participants tended to find more frequent words more challenging to define ($p < 0.1$). This result is difficult to interpret, however, a clue might be that more frequent words are typically encountered in a wider variety of contexts, and, thus, more difficult to define semantically.

Subtitles did not predict the 16-year-old group performance on the word definition task either. The reason why the English and the Norwegian subtitles group still appear in this table is due to the interaction between subtitles Condition and word frequency

Table 6 | Results of the inferential analysis of the word definition task (both age groups).

	Estimate	*SE*	*z*-value	Pr(>|z|)
(Intercept)	−19.230810	8.334339	−2.307	0.02103*
Age	0.588566	0.445330	1.322	0.18629
GroupEng	3.727448	5.028121	0.741	0.45850
GroupNor	10.754150	5.256792	2.046	0.04078*
Sexmale	0.266994	0.142786	1.870	0.06150(*)
log(FreqWD)	−0.413048	0.215268	−1.919	0.05502(*)
log(Grammar)	2.877757	0.954838	3.014	0.00258**
FilmSubeng	−0.202369	0.236279	−0.856	0.39173
FilmSubnor	−0.341372	0.146421	−2.331	0.01973*
Cartoon	−0.090802	0.052144	−1.741	0.08162(*)
Read_Norw	−0.002307	0.094375	−0.024	0.98050
Read_Eng	0.199213	0.129938	1.533	0.12524
Write_Eng	0.392285	0.126716	3.096	0.00196**
Listen_Eng	0.233386	0.131074	1.781	0.07498(*)
Total_Eng	−0.300474	0.184938	−1.625	0.10422
Age:GroupEng	−0.224494	0.306690	−0.732	0.46418
Age:GroupNor	−0.630797	0.319647	−1.973	0.04845*

*Age, Age group; GroupEng, English subtitles Condition; GroupNor, Norwegian subtitles Condition; sexmale, male sex; log (FreqWD), word frequency; log(Grammar), grammar test results (log); FilmSubeng, English subtitles when watching an English movie; FilmSubnor, Norwegian subtitles when watching an English movie; Cartoon, amount of time spent watching animated cartoons in English; Read_Norw, self-estimated Norwegian reading skills; Read_Eng, self-estimated English reading skills; Write_Eng, self-estimated English writing skills; Listen_Eng, composite factor created by the ratio of self-estimated English listening skills compared to self-estimated English speaking skills; Age: GroupEng, interaction of age and viewing condition with English subtitles; Age: GroupNor, interaction of age and viewing condition with Norwegian subtitles. Significance codes: (*) $p \leq 0.1$; * $p \leq 0.05$; ** $p \leq 0.01$.*

Table 7 | Results of the inferential analysis of the word definition task (17-year-old group).

	Estimate	*SE*	*z*-value	Pr(>|z|)
(Intercept)	−17.16533	7.04609	−2.436	0.01484*
log(Grammar)	5.02720	1.82148	2.760	0.00578**
log(FreqWD)	−0.57120	0.30673	−1.862	0.06257(*)
Listen	0.44393	0.26054	1.704	0.08840(*)
Eng_game	0.26633	0.06215	4.285	1.83e-05***

*log(grammar), grammar test results (log); log(FreqWD), frequency of the words (log); Listen, composite factor created by the ratio of self-estimated English listening skills compared to self-estimated English speaking skills; Eng_game, amount of time spent playing English computer games. Significance codes: (*) $p \leq 0.1$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.*

in this task. Frequency alone is not a significant predictor of the results across Conditions. The only effect in this respect is for participants in the Norwegian subtitle Condition, suggesting that this group performed better when the frequency of the word was higher ($p < 0.05$). The most significant factor, however, was a composite one created from the ratio of the scores on the vocabulary test and the participants' self-estimated English writing skills ($p < 0.001$). This factor, which most likely reflects vocabulary knowledge (including orthographic representations) best predicted performance on the word definition task.

Table 8 | Results of the inferential analysis of the word definition task (16-year-old group).

	Estimate	*SE*	*z*-value	Pr(>|z|)
(Intercept)	−10.12281	1.49018	−6.793	1.10e-11***
GroupEng	−0.12572	0.17645	−0.713	0.4762
GroupNor	0.28114	0.20975	1.340	0.1801
log(frequencyWD)	−0.13184	0.09394	−1.403	0.1605
log(vocabulary)	1.08642	0.14594	7.444	9.74e-14***
GroupEng:log(frequencyWD)	−0.01348	0.05028	−0.268	0.7886
GroupNor:log(frequencyWD)	0.11875	0.05847	2.031	0.0422*

*GroupEng, English subtitles Condition; GroupNor, Norwegian subtitles Condition; log(frequencyWD), frequency of the words (log); log(vocabulary), composite factor created by the ratio of the vocabulary test score compared to self-estimated English written skills (log); GroupEng:log(frequencyWD), word frequency with results from GroupEng; GroupNor:log(frequencyWD), word frequency with results from GroupNor. Significance codes: * $p \leq 0.05$; *** $p \leq 0.001$.*

Word recall task

Like in the word definition task, subtitles were not predictive of the participants' performance on the word recall task, suggesting again that there were no long term effects of the subtitles. Other predictors were, however, found. The results from the inferential analysis of the word recall task are showed in **Table 9**.

Table 9 shows that the participants who had spent more time watching English cartoons performed better on the word recall task ($p < 0.01$). The time they had spent writing English was also predictive of the performance ($p < 0.05$), indicating the significance of writing and orthographic representations for performance on this task.

DISCUSSION

SHORT TERM EFFECTS

Both the 16-year-old and the 17-year-old participants who watched the *Family Guy* episode with subtitles in either their L1 or their L2 performed better than the control groups on the comprehension questionnaire. This result emerged both in the combined and in the separate age-group analysis. Thus, positive short term effects of subtitles as a source of input and as a source facilitating processing of the authentic auditory input were found. Furthermore, for the 17-year-old group, the language of the subtitles did not seem to matter with both subtitling conditions having a similar effect ($p < 0.01$). Similar positive effects were found for the 16-year-old group, but, surprisingly, for that age group, the English subtitles were more facilitatory ($p < 0.01$) than the Norwegian subtitles ($p < 0.05$). These findings are in contrast to what we had expected, but support the views of Baltova (1999) who argues that the combination of auditory material in the target language (L2), verbal visual information, and nonverbal visual information in audio-visual

Table 9 | Results of the inferential analysis of the word recall task (both age groups).

	Estimate	*SE*	*z*-value	Pr(>|z|)
(Interce pt)	0.311592	0.341998	0.911	0.36225
Cartoon	−0.079072	0.028127	−2.811	0.00493**
Read_Norw	0.029137	0.051416	0.567	0.57093
Listen_Eng	0.005233	0.067037	0.078	0.93778
Total_Eng	0.141125	0.078840	1.790	0.07345(*)
Write_EngT	−0.086192	0.036128	−2.386	0.01705*

*Cartoon, amount of time spent watching animated cartoons in English; Read_Norw, self-estimated Norwegian reading skills; Listen_Eng, composite factor created by the ratio of self-estimated English listening skills compared to self-estimated English speaking skills; Write_EngT, how often the participants write English text. Significance codes: (*) $p \leq 0.1$; * $p \leq 0.05$; ** $p \leq 0.01$.*

material creates a better environment for learning than when only two of the three are available as input channels. When exposed to an animated cartoon in the L2, the students performed significantly better on comprehension of the contents of the episode when subtitles were available. Furthermore, we find evidence that target language subtitles enhance speech segmentation by providing the learner with orthographic information, in addition to the phonological one, and thus enhance the processing of the target language content (Mitterer and McQueen, 2009).

The combined age-group analysis revealed an important long-term effect of watching subtitled video material. This is reflected in the negative impact the habit of watching inter-lingual (native language) subtitles has on comprehension, as reported in the background questionnaire, and as seen in the predictive significance of this factor in the combined group analysis. A similar negative effect of frequent watching subtitled video with native language subtitles was also found in the combined analysis for the word definition task.

The results from both participant groups thus confirmed the overall initial hypothesis that having subtitles available would enhance the participants' performance, at least on the comprehension questionnaire. However, as it was hypothesized that the more advanced group would benefit more from authentic L2 subtitles as a result of higher proficiency, the lack of difference between the two subtitling Conditions in that group was surprising. Also, for the 16-year-old group it was assumed that native language (L1) subtitles would be more facilitative as a result of (somewhat) lower proficiency. However, the analysis revealed that the target language (L2) subtitles were more facilitative. One can argue that for the more advanced participants the language of the subtitles did not matter. What mattered was simply that they had access to a third input channel which assisted the comprehension of the contents of the audio-visual material.

The analysis of the 17-year-old group's performance can be argued to be contrary to the findings in Mitterer and McQueen (2009), Markham (1999), Vandergift (2011), and Vanderplank (1988), who found that L2 subtitles were more facilitatory. Mitterer and McQueen (2009) also argued that L1 subtitles would harm target language speech perception. However, our study indicates that native language subtitles may enhance comprehension of audio-visual material at least equally as target language subtitles. Minimally, these results argue against the view that the presence of native language subtitles harms the participants' perception of the auditory material. Bianchi and Ciabattoni (2008) found that target language subtitles were more facilitative for more advanced students, whereas native language subtitles were the better option for less advanced students. The results from our study, however, indicate that for more advanced students, the language of the subtitles is of a lesser importance, whereas for less advanced students, L2 subtitles were in fact more facilitatory. This is in line with Guillroy (1998), who argues that L2 subtitles can compensate for the challenging vocabulary in audio-visual material, and exactly this compensating effect of the English subtitles might have been a factor in boosting the performance of the 16-year-old participants in the English subtitles Condition. Moreover, we also found evidence that native language subtitles impact negatively on performance on comprehension from a long-term perspective, as revealed by the combined age-group analysis on both the comprehension task and the word definition task conducted 4 weeks after watching the test video.

An alternative explanation may be that the 16-year old participants in our study are advanced enough and are already at a high level of proficiency, e.g., compared to the participants in some of the studies reviewed here. This seems very likely in view of the advanced level of English competency in Norwegian school students overall (Alabau et al., 2002; Helland, 2008). Still, we find a difference between the age groups in the current study. We can interpret these results as consistent with earlier findings that target language subtitles are more facilitatory at high levels of proficiency, as reflected in the effects found in the 16-year-old group. It can then be speculated that the results of the 17-year-old group reflect a highly advanced proficiency level, where the language of the subtitles does not matter in skilled L2 comprehenders, especially on a single viewing instance (short-term). Furthermore, other factors, such as vocabulary size, grammar competence, daily L2 practices, such as watching target language subtitles and playing computer games are significant predictors of performance on the comprehension task, consistent with language learning research and the role of exposure to input (Mackey, 1999; Unsworth et al., 2014).

LONG TERM EFFECTS

In the combined analysis we found a long-term effect of the native language subtitle viewing Condition. However, this effect was modulated by age as seen in the interaction of age group and viewing condition. Furthermore, the separate analyses for each age group, revealed that subtitles were not predictive of performance on the word definition task or on the word recall task. As such, no long-term effects of the presence of subtitles on a single instance of viewing were found. Originally, we had expected to see an effect of the subtitles, and the tasks were designed directly based on the episode participants had watched 4 weeks earlier. One of the reasons for the lack of subtitling effects on these two tasks might be the long lapse between exposure to the stimuli and the testing. Four weeks is a long period of time and the episode was not watched again in this time period. Moreover, participants were

exposed to the material only once. Indeed, many studies which fail to document implicit learning in the context of authentic input have been criticized for testing too late and only after single exposure to the material tested. Since the subtitles did have an effect on the short term comprehension task, it is unreasonable to assume that participants did not pay attention to the subtitles, and that the absence of long-term effects on a single exposure was not caused by lack of attention to the subtitles. Moreover, there is ample evidence that subtitles and auditory material are processed in parallel, with subtitles automatically read (d'Ydewalle and Van de Poel, 1999) and that participants attend visually to the text (Bisson et al., 2014). Conducting the long term effect tasks sooner after exposure to the stimuli might have potentially led to different results, though the notion of "long term" thus becomes open to debate. Alternatively, exposing the participants to subtitled audio-visual material regularly during the 4 week period might also have led to positive implicit learning outcomes.

We do find, however, long-term effects of subtitle viewing in the combined age-group analysis based on self-reported daily practice of watching English films with either native language or target language subtitles. Our results indicate that a preference for watching native language subtitles has a negative impact on comprehension, as well as lexical skills, such as assessing word definitions, and thus, indirectly confirms the importance of target language subtitles (as compared to no subtitles at all). Finally, the word recall task only taps (most probably, conscious) memory of having encountered the target word in the context of the video, and as such, the results from this task are limited to interpretation. As noted by Vandergift (2011), the long term effects of subtitles are unclear, and we suggest that further research should be conducted in this area.

OTHER FACTORS

In this study we found mainly evidence of short term effects of the use of subtitles. However, other factors from the participants' linguistic background and L2 practices are worth discussing.

The combined analysis for the word definition task and the word recall task revealed a strong effect of overall self-assessed English writing skills, as well as English grammar competence on the word definition task only, as measured on the Grammar pretest. Both of these results suggest that overall L2 language competence, and specifically writing skills, impact on performance on a variety of lexical tasks and appear to underlie an important aspect of L2 lexical skills, most probably related to orthography and entrenching associations between form and meaning (Brown and Hulme, 1996).

For the 17-year-old group, the large effect of the amount of time spent playing English computer games ($p < 0.001$) on both the comprehension questionnaire and the word definition task is particularly interesting. Like films, computer games can be argued to provide authentic audio-visual material and to have an impact on the L2 acquisition of the learner. Computer games have similar visual features: they are animated and often contain both orthographic (textual) and auditory information. Thus students who are skilled players, are accustomed to interpreting meaning from animated L2 material, and might thus have an advantage on performance on comprehension in the cartoon task as well. Interacting with computer games in the L2 also requires learning new words in order to understand and proceed in the game, something which might have developed the player's heuristics skills as well as increased their vocabulary size. Indeed, there is a growing body of evidence suggesting incidental L2 learning in the context of computer games (Uzun et al., 2013; Sundqvist and Sylvén, 2014). The role of computer games can be accounted for within a situated and embodied cognition model of processing. On such a model, text comprehension is based on creating situation models that match the verbal content of the utterances (Kintsch, 1998; Zwaan, 1999). Computer games as a single factor influencing both participants' performance on the word definition task, and the comprehension task, is thus not so difficult to explain. Also, the effect of this factor can be mediated by (better) attentional skills, which can be tested in future work.

It is worthy to mention in the results for the 17-year-old group the predictive role of grammar competence on both comprehension and lexical skills (word definitions) ($p < 0.05$). One might argue that this is evidence that underlying grammar competence is an overall comprehensive predictor of performance on tasks in the L2, irrespective of their nature.

For the 16-year-old group, the results from the initial vocabulary test in the Norwegian subtitles Condition can be argued to be particularly interesting. In that group, vocabulary scores were an important factor both for comprehension ($p < 0.001$) and word knowledge (the word definition task) ($p < 0.001$), though for the latter it was a composite factor compared with self-estimated English writing skills. Indeed vocabulary knowledge has previously been established as a determinant of comprehension, particularly in the second language, though in younger participants (Lervåg and Aukrust, 2010). As the vocabulary test was taken before exposure to the material, we cannot view it as indicative of learning in this study. Instead, the results suggest that participants with better vocabulary knowledge (based on our testing) perform better on the comprehension test exactly as a result of a larger lexicon in the L2. The analysis of the result from the word definition task showed that the participants in the Norwegian subtitles Condition benefited less from the presence of subtitles than participants in the English subtitles Condition. A potential explanation can be sought in the Norwegian subtitles group's high scores on the vocabulary test, as this was generally more predictive of the results overall, thus highlighting the role of multiple factors in language acquisition. Furthermore, the analysis of the results from the comprehension task implies that participants in that group in all likelihood paid attention to the subtitles, still vocabulary knowledge superseded the importance of subtitles, suggesting that already obtained knowledge replaced the dependence on subtitles for higher competence learners.

We found that participants' self-assessed oral target language skills predict performance on the comprehension task. This result highlights the links between auditory perception skills in language learning and comprehension and indirectly support the idea that target language subtitles might be aiding in the process of L2 speech segmentation.

Worth mentioning is the contrast between results based on data from the self-reported L2 skills and L2-related activities. We found that viewing English video material (cartoons and

films) impacts negatively on performance on the comprehension task (combined age-group analysis). This is surprising given that exposure to target language input should matter and rather have a positive effect. This is also in contrast to the negative long-term effect of viewing native (not-target) language subtitles. However, watching can be a passive activity, and it is unclear whether participants attend to the audio (language) of the video or rather to the visual features. It has been suggested that interaction is more important than passive contexts, and especially L2 contexts require active interaction (Mackey, 1999; Oliver and Mackey, 2003). Moreover, we find an effect of playing computer games, which are interactive and involve motivation for moving on to a next level of the game, confirming the idea that language learning is best situated in interactive contexts. This is also consistent with findings in L1 acquisition research.

CONCLUSION

The question of whether second language learners should be trained using adapted materials or authentic materials is subject to debate. Many authors argue that learners should not be "protected" by adapted materials, and further suggest that authentic materials provide meaningful exposure to the target language (see Tomlinson, 2012 for a discussion). This study aimed to investigate the effects of subtitles in the context of authentic material (a cartoon video) on second language comprehension and potentially, second language acquisition for Norwegian learners of English. One hundred and fourteen participants in all participated in the study: 49 17-year-old students and 65 16-year-old students, who were all native speakers of Norwegian learning English as an L2 in high school. Both age groups were divided into three Conditions, where one group watched an episode of the American animated cartoon *Family Guy* with Norwegian subtitles, one group watched the episode with English subtitles, and one group watched the episode with no subtitles. On a comprehension questionnaire conducted immediately after watching the episode positive short term effects of both native language (L1) and target language (L2) subtitles were found for both age groups. However, no differences in terms of the language of the subtitles were found in the older and more advanced group. Four weeks later the participants responded to a word definition task and a word recall task to investigate potential long term effects of the subtitles. The only long-term effect of viewing subtitles on a single instance was found in the word definition task and was modulated by age. We found, however, that native language subtitles impact negatively on performance on the comprehension task. The results from this study suggest that the mere presence of subtitles as an additional source of information enhances learners' comprehension of the plot and content in animated audio-visual material in their L2. Since no major differences in terms of the language of the subtitles were found in the more advanced group, we argue that both intralanguage and interlanguage subtitles can aid target language comprehension in very advanced learners, most probably suggesting better consolidated vocabulary knowledge in that group. Furthermore, we found a difference between the two age groups in what best predicted performance on the two lexical tasks. While in the less proficient and younger group vocabulary status best predicted performance on both tasks (vocabulary predicts vocabulary), for the very advanced and older group, grammar was a stronger predictor, highlighting the importance of generic language competence and skills in L2 tasks for highly proficient L2 users. We also found an effect of written L2 skills on performance on both lexical tasks indicative of the role of orthography in vocabulary consolidation.

The current study has its limitations. We did not test participants' working memory skills or other cognitive competencies known to affect language acquisition and use (Vulchanova et al., 2014). We did not test attentional skills, which have been shown to have a bidirectional relationship with language competence and skills (as seen in studies of a bilingualism, Bak et al., 2014). Other behavioral measures, such as reaction times or eye-gaze data when viewing the experimental video could have added to assessing participants' performance. These are objectives for future research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01510/abstract

REFERENCES

Alabau, I., Bonnet, G., de Bot, K., Bramsbye, J., Dauphin, L., Erickson, G., et al. (2002). *The Assessment of Pupils' Skills in English in Eight European Countries*. Paris: European Network of Policy Makers for the Evaluation of Education Systems.

Bahrani, T., and Soltani, R. (2011). The pedagogical values of cartoons. *Res. Hum. Soc. Sci.* 1, 19–22.

Bak, T., Vega-Mendoza, M., and Sorace, A. (2014). Never too late? An advantage on tests of auditory attention extends to late bilinguals. *Front. Psychol.* 5:485. doi: 10.3389/fpsyg.2014.00485

Baltova, I. (1999). Multisensory language teaching in a multidimensional curriculum: the use of authentic bimodal video in core French. *Can. Mod. Lang. Rev.* 56, 31–48. doi: 10.3138/cmlr.56.1.31

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Benavent, G. T., and Peñamaría, S. S-R. (2011). Use of authentic materials in the ESP classroom. *Encuentro* 20, 89–94.

Bianchi, F., and Ciabattoni, T. (2008). "Captions and Subtitles in EFL Learning: an investigative study in a comprehensive computer environment", in *From Didactas to Ecolingua: an Ongoing Research Project on Translation and Corpus Linguistics*, eds A. Baldry, M. Pavesi, C. T. Torsello, and C. Taylor (Trieste: Edizioni Università), 69–90.

Bird, S. A., and Williams, J. N. (2002). The effect of bimodal input on implicit and explicit memory: an investigation into the benefits of within-language subtitling. *Appl. Psycholing.* 23, 509–533. doi: 10.1017/S0142716402004022

Bisson, M.-J., Van Heuven, W. J. B., Conklin, K., and Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: an eye tracking study. *Appl. Psycholing.* 35, 399–418. doi: 10.1017/S0142716412000434

Bolker, B., Brooks, M., Clark, C., Geange, S., Poulsen, J., Stevens, M., et al. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24, 127–135. doi: 10.1016/j.tree.2008.10.008

Brown, G., and Hulme, C. (1996). "Non-word repetitions, STM, and word age-of-acquisition: a computational model," in *Models of Short-term Memory*, ed S. E. Gathercole (Hove: Psychology Press), 129–148.

Chiquito, A. B. (1995). Metacognitive learning techniques in the user interface: advance organizers and captioning. *Comp. Hum.* 28, 211–223. doi: 10.1007/BF01830268

Dahl, A., and Vulchanova, M. (2014). Naturalistic acquisition in an early language classroom. Front. Psychol. 5:329. doi: 10.3389/fpsyg.2014.00329

Danan, M. (2004). Captioning and subtitling: undervalued language learning strategies. *Meta* 49, 67–77. doi: 10.7202/009021ar

Day, R. (2003). "Authenticity in the design and development of materials", in *Methodology and Materials Design in Language Teaching: Current Perceptions and Practices and their Implications*, ed W. A. Renandya (Singapore: SEAMEO Regional Language Centre), 1–11.

d'Ydewalle, G., and Gielen, I. (1992). "Attention allocation with overlapping sound, image, and text," in *Eye Movements and Visual Cognition: Scene Perception and Reading*, ed K. Rayner (New York, NY: Springer-Verlag), 415–427.

d'Ydewalle, G., and Van de Poel, M. (1999). Incidental foreign-language acquisition by children watching subtitled television programs. *J. Psycholing. Res.* 28, 227–244.

Ellis, N. C. (2013). "Second language acquisition," in *Oxford Handbook of Construction Grammar*, eds G. Trousdale and T. Hoffmann (Oxford: Oxford University Press), 365–378.

Flowerdew, J., and Peacock, M. (2011). *Research Perspectives on English for Academic Purposes.* Cambridge: Cambridge University Press.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Lang. Teach.* 40, 97–118. doi: 10.1017/S0261444807004144

Guillroy, H. G. (1998). The effects of keyword captions to authentic French video on learner comprehension. *Calico J.* 15, 89–108.

Gunderson, L., Odo, D. M., and D'Silva, R. (2011). "Second language literacy," in *Handbook of Research in Second Language Teaching and Learning,* Vol. 2, ed E. Hinkel (London: Routledge), 472–487.

Helland, T (2008). "Second language assessment in dyslexia: principles and practice, Chap. 3," in *Language Learners with Special Needs. An International Perspective*, eds J. Kormos and E. H. Kontra (Bristol: Multilingual Matters), 63–85.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge, UK: Cambridge University Press.

Lervåg, A., and Aukrust, V. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *J. Child Psychol. Psychiatry* 51, 612–620. doi: 10.1111/j.1469-7610.2009.02185.x

Mackey, A. (1999). Input, interaction and second language development: an empirical study of question formation in ESL. *Stud. Sec. Lang. Acquisit.* 21, 557–587. doi: 10.1017/S0272263199004027

Markham, P. (1999). Captioned videotapes and second language listening word recognition. *Foreign Lang. Ann.* 32, 321–328. doi: 10.1111/j.1944-9720.1999.tb01344.x

Mitterer, H., and McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS ONE* 4:e7785. doi: 10.1371/journal.pone.0007785

Neuman, S. B., and Koskinen, P. (1992). Captioned television as 'comprehensible input': effects of incidental word learning from context for language minority students. *Read. Res. Q.* 27, 95–106. doi: 10.2307/747835

Oliver, R., and Mackey, A. (2003). Interactional context and feedback in child ESL classrooms. *Mod. Lang. J.* 87, 519–543. doi: 10.1111/1540-4781.00205

Sherman, J. (2003). *Using Authentic Video in the Language Classroom.* Cambridge, UK: Cambridge University Press.

Sundqvist, P., and Sylvén, L. K. (2014). Language-related computer use: focus on young L2 English learners in Sweden. *ReCALL* 26, 3–20. doi: 10.1017/S0958344013000232

Tomlinson, B. (2012). Materials development for language learning and teaching. *Lang. Teach.* 45, 143–179. doi: 10.1017/S0261444811000528

Unsworth, S., Argyri, F., Cornips, L., Hulk, A., Sorace, A., and Tsimpli, I. (2014). The role of age of onset and input in early child bilingualism in Greek and Dutch. *Appl. Psycholing.* 35, 765–805. doi: 10.1017/S0142716412000574

Uzun, L., Cetinavci, U. R., Korkmaz, S., and Salihoglu, U. (2013). Developing and applying foreign language vocabulary learning and practicing game: the effect of vocaword. *Digit. Cult. Educ.* 5:1, 48–70.

Vandergift, L. (2011). "Second language learning: presage, process, product and pedagogy," in *Handbook of Research in Second Language Teaching and Learning,* Vol. 2, ed E. Hinkel (London: Routledge), 455–471.

Vanderplank, R. (1988). The value of teletext sub-titles in language learning. *ELT J.* 42, 272–281. doi: 10.1093/elt/42.4.272

Verspoor, M., Lowie, W., and de Bot, K. (2009). "Input and second language development from a dynamic perspective," in *Input Matters in SLA*, *1st Edn.*, eds T. Piske and M. Young-Scholten (Bristol: Multilingual Matters), 62–80.

Verspoor, M., and Winitz, H. (1997). Assessment of the lexical-input approach for intermediate language learners. *IRAL* 35, 61–75.

Vulchanova, M., Foyn, C., Nilsen, R., and Sigmundsson, H. (2014). Links between phonological memory, first language competence and second language competence in 10 year-old children. *Learn. Indiv. Diff.* 35, 87–95. doi: 10.1016/j.lindif.2014.07.016

Zarei, A. A., and Rashvand, Z. (2011). The effect of interlingual and intralingual, verbatim and nonverbatim subtitles on L2 vocabulary comprehension and production. *J. Lang. Teach. Res.* 2, 618–625. doi: 10.4304/jltr.2.3.618-625

Zwaan, R. A. (1999). Situation models: the mental leap into imagined words. *Curr. Dir. Psycholing. Sci.* 8, 15–18. doi: 10.1111/1467-8721.00004

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 April 2014; accepted: 08 December 2014; published online: 09 January 2015.

Citation: Vulchanova M, Aurstad LMG, Kvitnes IEN and Eshuis H (2015) As naturalistic as it gets: subtitles in the English classroom in Norway. Front. Psychol. ***5****:1510. doi: 10.3389/fpsyg.2014.01510*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 December 2014
doi: 10.3389/fpsyg.2014.01449

Raspberry, not a car: context predictability and a phonological advantage in early and late learners' processing of speech in noise

Kira Gor*

Graduate Program in Second Language Acquisition, School of Languages, Literatures, and Cultures, University of Maryland, College Park, MD, USA

Edited by:
Christos Pliatsikas, University of Kent, UK

Reviewed by:
Tamara Viktorovna Rathcke, University of Kent, UK
Angelos Lengeris, Aristotle University of Thessaloniki, Greece

****Correspondence:***
Kira Gor, Graduate Program in Second Language Acquisition, School of Languages, Literatures, and Cultures, University of Maryland, 3215 Jiménez Hall, College Park, MD 20742, USA
e-mail: kiragor@umd.edu

Second language learners perform worse than native speakers under adverse listening conditions, such as speech in noise (SPIN). No data are available on heritage language speakers' (early naturalistic interrupted learners') ability to perceive SPIN. The current study fills this gap and investigates the perception of Russian speech in multi-talker babble noise by the matched groups of high- and low-proficiency heritage speakers (HSs) and late second language learners of Russian who were native speakers of English. The study includes a control group of Russian native speakers. It manipulates the noise level (high and low), and context cloze probability (high and low). The results of the SPIN task are compared to the tasks testing the control of phonology, AXB discrimination and picture-word discrimination, and lexical knowledge, a word translation task, in the same participants. The increased phonological sensitivity of HSs interacted with their ability to rely on top–down processing in sentence integration, use contextual cues, and build expectancies in the high-noise/high-context condition in a bootstrapping fashion. HSs outperformed oral proficiency-matched late second language learners on SPIN task and two tests of phonological sensitivity. The outcomes of the SPIN experiment support both the early naturalistic advantage and the role of proficiency in HSs. HSs' ability to take advantage of the high-predictability context in the high-noise condition was mitigated by their level of proficiency. Only high-proficiency HSs, but not any other non-native group, took advantage of the high-predictability context that became available with better phonological processing skills in high-noise. The study thus confirms high-proficiency (but not low-proficiency) HSs' nativelike ability to combine bottom–up and top–down cues in processing SPIN.

Keywords: heritage language speakers, speech in noise, early and late learners, second language acquisition, language proficiency, non-native speech recognition, context predictability, phonological sensitivity

INTRODUCTION

WHO ARE HERITAGE SPEAKERS?

More people in the world are raised bilingual or multilingual than monolingual (Bhatia and Ritchie, 2013, XXI). Among the millions of bilingual speakers across the world, there is a group that have been called heritage speakers (HSs). HSs are early interrupted learners, who acquire their first language naturalistically as infants at home from their caregivers, but who switch to the language spoken in the community in their childhood (Valdés, 2005; Polinsky, 2008). As a result, second language (L2) becomes the dominant language of HSs, and their first (L1), heritage, language is reduced to non-native levels of proficiency due to incomplete acquisition and/or attrition (Montrul, 2008; Bylund, 2009; Bylund et al., 2010; Schmid, 2010; Polinsky, 2011). The heritage language may also be influenced by L2, the dominant language (Cook, 2003; Polinsky, 2014). HSs rely predominantly on auditory input, and often do not go through formal schooling in their first language. Due to this auditory bias, they typically prefer the listening and speaking modalities, have poor reading and writing skills, and are sometimes illiterate. HSs, early starters with non-native proficiency in their first language, have recently attracted the attention of researchers. And indeed, understanding the role of early start (from birth) in shaping the linguistic profile and the underlying processing mechanisms of HSs as opposed to late L2 starters makes it possible to address the critical period hypothesis (Abrahamsson and Hyltenstam, 2009; Bylund et al., 2012; DeKeyser, 2013). At the same time, HSs are compared to native speakers since both populations acquire language naturalistically from birth. This allows researchers to identify native and non-native aspects of heritage language (Montrul, 2012), and to establish the role of incomplete acquisition as opposed to attrition (Bylund et al., 2010).

Late L2 learners, unlike heritage language speakers, start learning their second language as adults, after puberty. The type of L2 exposure, naturalistic or formal classroom, depends on biographic trajectories of individual L2 learners, and on global migration patterns for larger populations of learners. Demographic trends, including the patterns of migration, often determine which populations of L2 learners will study L2 in a foreign language classroom, and which will actually move to the country where L2 is spoken. Formal late L2 learners, and university students in particular, often rely heavily on visual input (Psaltou-Joyceya and Kantaridoub,

2011). While there exists a range of methodologies for teaching a foreign language to late learners in a classroom setting outside the target language community, university-level academic programs in the U.S. typically introduce reading in Russian from the outset (Gor, 2000). A perusal of the major Russian language textbooks for beginners currently used in American universities shows that they rely on reading from day 1 (Lubensky et al., 2002; Lekic et al., 2008; Robin et al., 2014). In this study, native speakers of American English and late L2 learners of Russian were all predominantly shaped by in-class experience, which could be complemented by an immersion. No late L2 learner in the sample was a naturalistic learner. Conversely, HSs acquire their heritage language from birth in a uniquely auditory modality. Research on HSs in comparison with adult native speakers and late L2 learners makes it possible to gage the role of early naturalistic exposure in shaping the mechanisms underlying auditory speech processing. The uniqueness of HSs lies in the fact that they have received early naturalistic input in the same way as native speakers, yet have reduced, non-native proficiency in their L1, and thus can be compared to late L2 learners at the same proficiency level to single out the influence of early naturalistic exposure and input.

To summarize, heritage language is a native language acquired naturalistically from birth from caregivers that does not reach native proficiency levels due to a switch to another language spoken in the community, which becomes the dominant language. Heritage languages are often spoken languages due to the reduced amount of schooling that heritage language speakers receive. While there is a growing number of studies addressing the domains of heritage language phonology (Oh et al., 2003; Chang et al., 2011; Lukyanchenko and Gor, 2011), morphology (Gor et al., 2009; Gor and Cook, 2010), morphosyntax (Montrul et al., 2008, 2013, 2014; Montrul, 2009, 2011), and syntax (Keating et al., 2011; Lee-Ellis, 2011; Polinsky, 2011), there have been no studies, to the best of our knowledge, exploring the robustness of heritage auditory sentence processing, and in particular, HSs' ability to rely on context predictability in adverse conditions, such as speech in noise (SPIN).

SPEECH IN NOISE AND TOP–DOWN AND BOTTOM–UP PROCESSING

Given that SPIN, as one of the adverse conditions, has been used to study the properties of the human speech recognizer (Mattys et al., 2012), it can become a powerful diagnostic tool for the robustness of non-native speech perception. Moreover, recent renewed interest in speech processing in adverse conditions, including different kinds of noise, stems from the understanding that (1) adverse conditions are ecologically more valid than unrealistic idealized listening conditions, e.g., clear speech (see Mattys et al., 2012), and (2) by manipulating the properties of noise and the listening materials, one gains insights into the complex interaction of top–down and bottom–up processing in different groups of listeners. Was it raspberry or car ('malina' or 'mashina,' correspondingly, in Russian) that was mentioned in the sentence? In noisy conditions, these two feminine nouns can be confused easily. However, the context in which they were heard usually disambiguates the word in question. The high cloze probability context, if recovered from noise, will disambiguate *car* and *raspberry* in Russian sentences 1a and 1b.

(1a) Okolo doma stojala staraja mashina.
Near house stood old car.NOM.SG.
'An old car stood near the house.'

(1b) V sadu rosla spelaja malina.
In garden grew ripe raspberry.NOM. SG.
'Ripe raspberries grew in the garden.'

Critically, the whole sentence is masked by noise, and not just the last word, and the listener therefore needs to recover sentence cues from the acoustically degraded signal. This means that the mechanisms of prediction and sentence integration need to rely on acoustic cues that are less than robust, starting from the beginning of the sentence and building up expectations by the last word. Note that Russian allows scrambling, but crucially, the word order with the sentence-final noun-subject is canonical for this particular sentence structure, with the adverbial phrase fronted. Context predictability was manipulated in the original SPIN test developed for native speakers of English (Kalikow et al., 1977) and later adapted for Spanish (Cervera and González-Alvarez, 2011). The role of prediction and its interaction with heritage and late L2 learner profiles and high/low-proficiency levels is the main focus of the present study.

NOISE TYPES AND INFORMATIONAL AND ENERGETIC MASKING

Before we address non-native processing of SPIN, let us revisit the understanding of the impact of different types of environmental degradation, including noise, on speech processing in native speakers. This will assist us in situating the present study and later in interpreting the findings with regard to the type of the noise that it used. There are two types of environmental degradation that are used in psycholinguistic experiments: energetic masking and informational masking (Van Engen and Bradlow, 2007; see Mattys et al., 2012 for a review). Energetic masking is created by the use of white noise or filtering and requires signal separation and lower-level acoustic encoding and activation of lexical-semantic information. Conversely, informational masking such as babble noise or speech compression interferes with higher-order selection and integration (Aydelott and Bates, 2004). The study by Aydelott and Bates (2004) used two types of distortion, low-pass filtering and 50% speech compression, and three types of priming sentence context, congruent, incongruent, and neutral. The format of the experiment was a lexical decision task with priming, where the priming context was manipulated, and the target final word (or non-word) was presented without distortions. The study recorded reduced facilitation in congruent low-pass filtered sentences, and reduced inhibition in incongruent compressed sentences compared to the neutral context. It concluded that energetic masking induced by low-pass filtering interfered with early low-level acoustic encoding and the activation of lexical entries, while sentence compression affected central language processing and sentence integration. While, there are no data at present on the impact of different adverse conditions on HSs' speech recognition, it is reasonable to assume that the involvement of different levels of speech processing depending on the type of distortion will be same as for native speakers.

The present study used a multi-talker babble noise, which sounds like the noise of many people talking at the same time

in the background. This type of noise is ecologically valid given its pervasive presence in everyday life. Note that listening to speech in adverse conditions is considered to be part of a listener's daily auditory experience rather than an extraordinary situation, and consequently, Mattys et al. (2012, p. 963) maintain that speech recognition in adverse conditions is synonymous with speech recognition *per se*. Thus, SPIN tests the robustness of non-native listeners' speech recognition under ecologically valid conditions.

Multi-talker babble noise combines both energetic and informational masking and thereby has a double effect on speech intelligibility. The superposition of several speech recordings on the target sentence produced a white noise component that is associated with energetic masking (Mattys et al., 2012). Energetic masking, as well as low-pass filtering, primarily affects the acoustic-phonetic properties of speech, and decreases its intelligibility by interfering with low-level processing. The more talkers, the more energetic masking takes place. At the same time, once the informational masking effect is partialled out, babble noise also produces informational masking that has different implications for speech intelligibility. Informational masking has higher-level consequences, as it leads to attentional capture, semantic interference, and eventually, increases the cognitive load. In the present study, the multi-talker babble had a high component of steady noise, but it also had an informational masking component, with a more limited competition between the informational streams than in a two-talker babble.

SPEECH IN NOISE IN NON-NATIVE PERCEPTION

There exists a large body of evidence that L2 speakers' perception of L2 speech in noisy conditions deteriorates to a greater extent than does the perception of native speakers (Kalikow et al., 1977; Mayo et al., 1997; Munro, 1998; van Wijngaarden et al., 2002). This effect has possible explanations involving redundancy reduction or fuzziness in L2 perception at different levels, from phonetic (e.g., uncertainty about phonetic contrasts) to semantic. Apparently L2 speakers do not make efficient use of the probabilities that context provides. "The levels of noise at which the speech was intelligible were significantly higher and the benefit from context was significantly greater for monolinguals . . . than for late bilinguals" (Mayo et al., 1997, p. 686).

While there is numerous evidence that non-native speech perception is affected by noisy conditions to a greater extent than native perception, there is no agreement regarding the relative role of several factors implicated in L2 learners' perceptual problems when processing SPIN. Reduced speech discriminability in SPIN has been demonstrated in L2 listeners for non-word syllables (Cutler et al., 2004, 2008; Rogers et al., 2006; Broersma and Scharenborg, 2010), isolated words presented in lists (Rogers et al., 2006), words embedded in a sentence (Mayo et al., 1997; Bradlow and Alexander, 2007; Oliver et al., 2012), and whole sentences (Meador et al., 2000; Bradlow and Bent, 2002; Pinet et al., 2011). Studies focusing on the role of different aspects of non-native speech processing affected by noise fall mainly into three categories. The first category focuses on sublexical processing of isolated phonemes, e.g., individual phonemic confusions for English intervocalic consonants (Garcia Lecumberri and Cooke, 2006; Cutler et al., 2008; Broersma and Scharenborg, 2010). The second category is concerned with the phonological/lexical interface and phonemic confusions associated with word recognition (Oyama, 1982; Meador et al., 2000; Cooke et al., 2008). And finally, the third explores the reliance on sentence context and the use of cloze probabilities (van Wijngaarden et al., 2002; Bradlow and Alexander, 2007). The priming role of the context presented in noise in native and non-native populations has been explored for word priming (Golestani et al., 2009, 2013; Hervais-Adelman et al., 2014), and sentence priming (Aydelott and Bates, 2004). Crucially, two studies exploring the behavioral and neural bases of semantic context use in word and sentence priming, showed a consistent semantic context advantage for native speakers, but not second language learners (Golestani et al., 2009, 2013; Hervais-Adelman et al., 2014).

Studies explore the use of sentence context and cloze probabilities in various ways. The SPIN test (Kalikow et al., 1977) compared recognition of the sentence-final word, with the preceding context either making the word highly probable or impossible to predict. Thus, if at least part of the sentence can be auditorily recovered from noise in 'The mouse was caught in the trap,' the listener is unlikely to hear 'tram' instead of 'trap.' At the same time, when the context does not support the choice of one word over the other, confusion is more likely to occur. In: 'They hope he heard about the rent,' the low cloze probability does not support either the actual or the alternative word, for example, 'tent.' A more radical approach to cloze probabilities was adopted by Meador et al. (2000) who created sentences with low transitional probabilities between each word in the sentence and the following one, as in: 'The blonde dentist ate the heavy bread.' There, participant's task was to repeat the sentence verbatim, and the accuracy score referred to the number of words that were correctly recovered from the sentence. The present study uses the approach of Kalikow et al. (1977), with two types of sentences differing by the probability of the last word only, which makes it possible to control for the properties of sentence-final words recognized in noise.

A study by Meador et al. (2000) directly addressed the relative role of non-native phonology in non-native word recognition in sentences. The study hypothesized that the native Italian participants' accuracy in perceiving English vowels and consonants would be related to their recognition of English words in sentences with low transitional probabilities between words, as in the example above. To verify this hypothesis, the authors regressed the segmental perception scores obtained for the native Italian participants in two other studies onto the word recognition scores, i.e., the number of repeated words in the sentence. The results support the role of phonological deficits (non-native consonant perception in that specific case) in SPIN recognition. However, the findings of the study are not sufficient to evaluate the role of non-native phonological perception as opposed to top–down use of context predictability, since the sentences used in the study had the lowest cloze probabilities possible.

No data are yet available on heritage processing of SPIN. Is SPIN perception in HSs on the par with native speakers because they have the advantage of early starters, or is it degraded as in L2 learners because their proficiency is comparable to late L2 learners? While there is robust evidence that non-native speech

perception is affected by noisy conditions to a greater extent than native perception, there is no agreement regarding the relative role of several factors implicated in L2 learners' perceptual problems when processing SPIN. These factors include phonological deficits, reduced lexical knowledge, and a reduced ability to rely on top–down processing and to use contextual cues for sentence integration. The current study fills the gap and compares the perception of Russian speech in multi-talker babble noise in HSs of Russian and late L2 learners at the same proficiency levels to that of native Russian speakers. HSs of Russian in the study are early interrupted learners whose first language spoken at home was Russian, but who later switched to English, currently their dominant language. Given that heritage language is shaped by early naturalistic exposure from birth that relies exclusively on the aural modality, at least in the first years of life, one can hypothesize that HSs would have a processing advantage for SPIN over late L2 learners. Indeed, late learners, college-level students, mainly acquire Russian in a formal classroom and rely heavily on visual input, i.e., reading. While the goal of a modern foreign language classroom is to develop all four skills—two receptive, reading and listening, and two productive, speaking and writing (Rogers, 2014)—an objective assessment of the listening skills in late learners of Russian as a foreign language produced disappointing results (Thompson, 2000, p. 276). If a heritage SPIN advantage were to be found, the question arises as to the factors underlying this advantage.

THE CURRENT STUDY

This study investigates the role of sentence context predictability and uses two levels of multi-talker babble noise, high and low, to determine whether the efficiency of processing SPIN depends on bottom–up acoustic-phonetic and/or top–down semantic-syntactic sentence integration. It goes on to compare the outcomes of the SPIN test with three additional tests of phonological and lexical knowledge in the same groups of participants[1]. To control for the role of possible phonological deficits leading to problems with efficient processing of acoustically degraded speech, the study uses two independent measures of phonological perception. Both measures target the phonological contrast that causes most difficulties for speakers of English, the hard/soft consonant contrast. The AXB discrimination task measures sensitivity to the contrast in nonsense syllables, while the picture-word discrimination task looks at the sensitivity to the same contrasts in minimal pairs of lexical items and thus investigates the robustness of phonolexical representations differentiated by the same hard/soft contrast. In order to explore the possibility that the advantage on the SPIN task may stem from superior knowledge of vocabulary, the study compares the accuracy scores on a multiple-choice task measuring vocabulary in different frequency ranges.

The study addresses the following questions:

- Are HSs as efficient as L1 speakers in listening to SPIN or do they experience the same deficits as late L2 learners at the same proficiency levels?
- Which factors are responsible for the problems experienced by HSs and L2 learners when processing SPIN: phonological deficits, lack of vocabulary knowledge, and/or the ability to rely on top–down processing and use sentence cues?
- What is the role of proficiency and learning background, early versus late start in the ability to rely on top–down processing?

EXPERIMENT 1: SPEECH IN NOISE

MATERIAL AND METHODS

The present study uses the design of the original SPIN test (Kalikow et al., 1977), with high- and low-probability sentences presented in two levels of noise, high and low, and the task for the participant was to repeat the last word of the sentence. It used balanced lists of words created based on a comprehensive study of Russian speech recognition in white noise, that has identified numerous factors that influence speech comprehensibility in both native and non-native speakers (Shtern, 1992). These factors form a hierarchical structure and depend on the type of stimuli: syllables, words, sentences, and extended text. Since the task in the current experiment elicits the responses at the word level, only the findings about this level are provided below. Shtern obtained the following hierarchy of factors at the word level (words presented in isolation) in native speakers that are relevant to the present study:

1. Length of the word in phonemes: the longer the word, the better it is perceived.
2. Part of speech: nouns are best, and verbs worst, in intelligibility.
3. Stressed vowel: /a-o-e-i/ have better intelligibility than /u-ɨ/[2].
4. Consonantal load: the more consonants in a word, the better its perception.
5. Place of stress: disyllabic words with stress on the first syllable are perceived better than those with stress on the second syllable.

The same study emphasized that the level of predictability, defined and measured by the presence and number of key words suggesting the use of the target word, plays an important role in speech intelligibility at the sentence level and above and interacts with the level of noise and purely phonetic factors described above at the word level. Shtern (2001) created balanced word lists in such a way, that each list of 10 nouns in the Nominative case (the citation form in Russian) has the same parameters that have been demonstrated to be critical for recognition of SPIN by native Russian speakers. The lists of nouns created by Shtern and used in this study are balanced in frequency (with four gradations; only relatively high-frequency words are used), length in syllables (two monosyllabic, four disyllabic, and four trisyllabic words), stress placement, stressed vowel (two of each vowel: /a/, /u/, /e/, and /o/, and one of each: /ɨ/ and /i/), and the percentage of voiceless consonants (40–50% per list). We used eight lists with 10 nouns each to create 80 sentences.

[1]The experiments reported in this publication are part of a larger research project Linguistic Correlates of Proficiency sponsored by the Center for Advanced Study of Language at the University of Maryland (see Long et al., 2012).

[2]Here and elsewhere in the text, /ɨ/ refers to the high central (or, more exactly, mixed) unrounded vowel that occurs after hard consonants in Russian. Its phonological status is controversial given that it is in complementary distribution with the vowel /i/ that occurs after soft consonants. The present article follows the position of the Leningrad/Saint Petersburg Phonological School (see Bondarko, 2005) and treats /ɨ/ as a separate phoneme.

Materials

The critical design of the SPIN used in this study crosses two factors: noise level and predictability of the final word based on the sentence context. In general, it is expected that higher noise levels will produce more errors. However, as proficiency increases, learners' perception should be more robust in the face of noise, because of a greater internalization of syntactic structure, semantic properties, collocational tendencies, phonological information, etc. Therefore, sentence context was manipulated to be either highly predictive of the final word (e.g., 'I don't have a sister, but I have a brother'), or not at all predictive (e.g., 'The man in the park has a brother'). It is expected that under very noisy conditions, advanced and near-native learners will show a large effect of context, where the words in highly predictive sentences are easier than the words in poorly predictive sentences. It is expected that this advantage of context will correlate with proficiency.

The task uses four conditions, with two levels of noise and two levels of context cloze probability. The high-noise level is combined with 20 high and 20 low cloze probability sentences. Identically, the low-noise level is combined with 20 high and 20 low cloze probability sentences. Thus, the task includes eight blocks of 10 sentences each—four high-probability (40 sentences), and four low-probability (40 sentences). The target word is a sentence-final noun. For the sentence-final word, the task uses phonetically balanced lists of nouns (Shtern, 2001). The carrier sentences, both high- and low-probability, were balanced for number of words (average 4.8 to 5.4 words depending on the block), and number of syllables (10.03 to 10.12 syllables). A total of 80 sentences were used. All participants listened to the same set of sentences, which made it possible to reduce the number of participants in the study and to ensure that no uneven distribution of participants with varying proficiency across different presentation lists takes place. This was imperative given that heritage and L2 participants were in the same proficiency range based on the standardized test of oral proficiency (see Participants). Sample items (2a,b) are provided below:

(2a) High cloze probability context

U	menja	net	sestry,	no	est'	brat.
At	me	no	sister	but	(there) is	brother.

'I don't have a sister, but I have a brother.'

(2b) Low cloze probability context

Rebjonok	ne	znal,	chto	eto	otvet.
Child	not	knew	that	this (was)	answer.

'The child did not know that this was the answer.'

Two voices, male and female, were used to record the stimulus sentences. Half of the sentences (40) were presented in the male voice, and another half in the female. Voices were not alternated, but presented in two blocks, first the male and then the female. The recordings were rescaled so that they had similar energy values. The multi-talker babble noise was produced by forward-superimposing multiple stimulus sentences from the same task so that the noise had a speech-shaped quality and the same frequency spectrum as the stimulus sentences. The level of the resulting noise was manipulated to create two noise conditions: low-noise and high-noise. The sentences were then combined with each of the two masker noise types such that the noise signal started on average 1.5 s before the onset of the sentence and continued for about 1.5 s after the sentence offset. The speech-to-noise ratio (SNR) for the low-noise condition was on average 4 dB, and the SNR for the high-noise condition was on average 1.5 dB. To determine the appropriate SNR for each sentence in the high- and low-noise conditions, a subjective piloting was used with four native speakers of Russian who did not take part in the experiment. Only sentences with the low-predictability context were used to establish the target noise level. In the high-noise condition, half of native listeners identified the last word in the sentence, while in the low-noise condition, three out of four did. Thus, the choice of the SNR for both noise conditions reflected average discriminability by native speakers of Russian established prior to the main experiment.

Participants

Sixty-eight people participated in the SPIN experiment and were paid for their participation. Specifically, the data were collected from 11 native speakers of Russian, 23 HSs, and 34 late L2 learners of Russian. The sample contained 31 males and 37 females. As seen in **Table 1**, the average age of the L2-high group is higher than that of the other participants, and L2 learners tend to be older on average. This tendency is understandable given that it takes several years to reach the low-level Russian proficiency threshold established in this study, and even longer to achieve very high proficiency. Given that the experiment did not collect reaction time data, these age differences are not expected to bias the results. The SPIN test was part of a larger 4-h long test battery (Gor and Cook, 2010; Long et al., 2012), and the results of the SPIN test are compared below to the tests gaging phonological discrimination and vocabulary control in the same heritage and L2 participant groups. HSs who participated in this experiment had Russian-speaking parents, were exposed to Russian from birth and heard it spoken at home on a daily basis. However, they had lived in the U.S. since the age of 7 on average (range: 0–14), and considered English to be their dominant language, and Russian, the language of the test, their weaker language. HSs did not live in Russia or a Russian-speaking country after puberty, and had little or no formal elementary schooling in the Russian language, although

Table 1 | Background information of the participants in the study.

Participant group	*N*	Age mean (range)	Gender M/F
Native speakers	11	25.55 (22–30)	3/8
Heritage speakers, high proficiency	12	24.08 (18–51)	4/8
Heritage speakers, low proficiency	11	20.81 (18–25)	3/8
L2 learners, high proficiency	18	41 (25–56)	12/6
L2 learners, low proficiency	16	28 (21–44)	8/8
Total	68	29.31 (18–56)	31/37

High proficiency includes oral proficiency levels 2+, 3, 3+, 4; low proficiency includes oral proficiency levels 1, 1+, 2 based on the Interagency Language Roundtable (ILR) oral proficiency scale.

all of them could read in Russian. Late L2 learners were all native speakers of American English and started learning Russian after puberty in a formal classroom, most of them as young adults in college. The average age of onset of Russian was 18.4 years (range: 13–27), and an average length of formal study was 10 years (range: 0–39). While all but five L2 learners had a study abroad experience in Russia or a Russian-speaking country, they did not learn Russian in a naturalistic setting, merely by virtue of living in a Russian-speaking country or community.

Heritage speakers and L2 learners of Russian in this experiment were divided into two groups, high- and low-proficiency, using the Interagency Language Roundtable (ILR) testing format, which made possible direct comparisons of the high- and low-level proficiency heritage and L2 participants (Long et al., 2012)[3]. The ILR score is established based on an audio-recorded oral proficiency interview conducted with a certified tester. The interview lasts 20–30 min and takes the form of a rigidly structured conversation, although the topics of the conversation vary depending on the testee's background. The ILR oral proficiency score is a standard global language proficiency score widely accepted in the U.S. In addition to the base levels, the ILR scale has "plus" sublevels that refer to the proficiency exceeding the requirements of the level. In our participant groups, both heritage and L2, the oral proficiency scores ranged from 1 (Intermediate) to 2 (Advanced), 3 (Superior), and 4 (Distinguished). Both the heritage and L2 samples also included "plus" sublevels, e.g., 1+ (Intermediate High). The participants were divided into low-proficiency groups containing participants with the ILR scores ranging from 1 to 2 (16 L2 and 11 HSs), and high-proficiency groups containing participants with ILR scores ranging from 2+ to 4 (18 L2 and 12 HSs). A detailed breakdown by age, gender, and proficiency level is provided in **Table 1**.

Procedure

The listening materials in the SPIN task were presented in two blocks of 40 sentences, the first recorded in a male voice and the second in a female voice, with a short pause between the blocks. Each set of 40 sentences included all four critical conditions, high-noise/high-predictability context, high-noise/low-predictability context, low-noise/high-predictability context, and low-noise/low-predictability context. The order of the sentences in these four conditions was randomized within each block (male-voice and female-voice), and was the same for all participants. Participants were tested individually, and were seated in a quiet room in front of *Dell® Latitude/D820* computers with *Plantronics.* Audio 750 headsets with mounted microphones and *Logitech® Precision USB* game pads. They were presented with instructions on the computer screen in English, and used buttons on their gamepad to initiate the following trial. Participants listened to the entire sentence in noise and were then asked to repeat the sentence-final word into the microphone. The experiment was self-paced and took ∼20 min. Participants were encouraged to take a break in the middle. All four experiments reported in the present publication were part of a larger test battery and were completed on the same day. Ample rest time was provided to participants to reduce possible fatigue. Also, the type of activity varied from one task to the next, which lessened the effect of monotony. The experiment was programmed in DMDX (Forster and Forster, 2003). Responses were recorded and then manually transcribed by trained linguists, native speakers of Russian. No substitutions were accepted when scoring the responses for accuracy. Only correct responses were scored as 1; all the other responses, e.g., responses with a phonological neighbor, were coded as 0. The accuracy score results were subjected to statistical analyses.

[3]The information about the Interagency Language Roundtable proficiency scale and the testing format can be found at: http://www.govtilr.org/skills/ILRscale1.htm

RESULTS

The accuracy scores for each participant group broken down by the level of noise and context predictability are presented in **Table 2** and **Figure 1**. Participants' responses were analyzed with a repeated measures ANOVA in by-subject and by-item analyses. The study had a 2 × 2 × 5 factorial design, with the following predictor variables: context predictability (two levels: low and high), noise level (two levels: low and high), and language proficiency group (five levels: L2-low, L2-high, HS-low, HS-high, Native). The dependent variable was the accuracy of correctly identified words in a

Table 2 | Participants' mean accuracy scores across all conditions.

	Language group				
	L2-low	**L2-high**	**HS-low**	**HS-high**	**Native**
HN/HC	0.36	0.44	0.51	0.64	0.85
HN/LC	0.22	0.23	0.30	0.28	0.45
LN/HC	0.84	0.93	0.95	0.94	0.98
LN/LC	0.68	0.73	0.74	0.79	0.79

HN, high noise; LN, low noise; HC, high context predictability; LC, low context predictability.

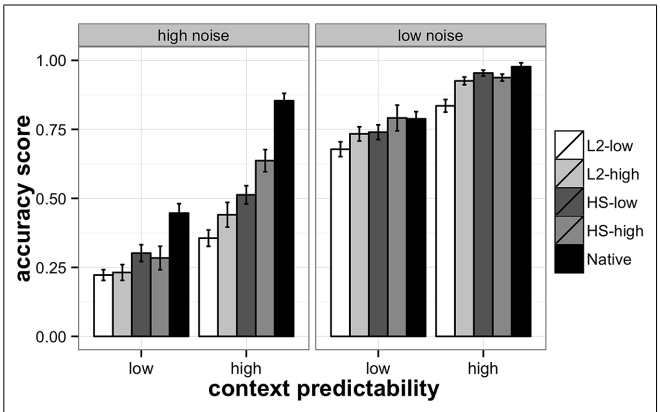


FIGURE 1 | Accuracy scores on SPIN task in heritage, L2, and native participants. Heritage and L2 participants are divided into high- and low-proficiency groups. The left panel represents the high-noise and the right the low-noise conditions. L2-low – low-proficiency L2 learners, L2-high – high-proficiency L2 learners, HS-low – low-proficiency heritage speakers, HS-high – high-proficiency heritage speakers, and Native – native speakers of Russian.

sentence. R statistical package was used for the analyses (R Core Team, 2013, version R 3.01). The results are represented in **Table 3** (by-subject) and **Table 4** (by-item).

The analysis revealed a significant context effect indicating that participants on average performed better in the high-predictability context condition. A significant noise effect suggests that word identification was significantly more accurate in the low-noise condition, and a language group effect supports the differences among the participant groups. There were also significant context by noise, context by group, and noise by group two-way interactions. Finally, a three-way interaction between context, noise and language group was also found significant, suggesting that the interaction between noise and context changed across the levels of the language group variable. Separate ANOVAs for each group showed that two-way interactions between context and noise were significant in the Native [$F_1(1,252) = 10.93$, $p < 0.01$; $F_2(1,2215) = 3.96$, $p = 0.05$], and the HS-high [$F_1(1,252) = 10.64$, $p < 0.01$; $F_2(1,2215) = 9.85$, $p = 0.01$], groups, while they were statistically insignificant in the L2-low, L2-high, and HS-low groups.

These data are represented visually in **Figure 2** where the difference between the accuracy score in the high-predictability and low-predictability conditions is provided as a percentage. This difference accounts for the context effects on response accuracy under the same noise levels.

Figure 2 demonstrates that while L2-low, L2-high, and HS-low groups benefited from high context predictability to a similar extent regardless of the noise condition (low or high), Native and HS-high groups appear to rely on context to a greater extent (almost 40% more) when they listen to sentences in high-noise compared to low-noise, or in other words, they take advantage of the context when it is both needed and available. TukeyHSD *post hoc* tests showed that the increasing group differences (from L2-low to Native group) in the accuracy scores in the high-noise/high

Table 3 | Repeated measures ANOVA results for Experiment 1: speech in noise, by-subject analyses.

By-subject

	Df	Sum Sq	Mean Sq	*F* value	Pr(>F)
Between-subject					
Group	4	1.767	0.4417	22.29	<0.000
Residuals	63	1.249	0.0198		
Within-subject					
Context type	1	3.089	3.089	322.633	<0.000
Noise	1	12.013	12.013	1254.523	<0.000
Context type:Noise	1	0.082	0.082	8.61	0.00376
Context type:Group	4	0.176	0.044	4.585	0.00148
Noise:Group	4	0.525	0.131	13.707	<0.000
Context type:Noise:Group	4	0.18	0.045	4.7	0.00122
Residuals	189	1.81	0.01		

Table 4 | Repeated measures ANOVA results for Experiment 1: speech in noise, by-item analyses.

By-item

	Df	Sum Sq	Mean Sq	*F* value	Pr(>F)
Between-item					
Context type	1	9.748	9.748	31.9	<0.000
Residuals	78	23.833	0.306		
Within-item					
Noise	1	32.95	32.95	1044.759	<0.000
Group	4	5.23	1.31	41.465	<0.000
Context type:Noise	1	0.22	0.22	7.087	0.00794
Context type:Group	4	0.58	0.15	4.602	0.00113
Noise:Group	4	1.71	0.43	13.584	<0.000
Context type:Noise:Group	4	0.44	0.11	3.509	0.00756
Residuals	702	22.14	0.03		

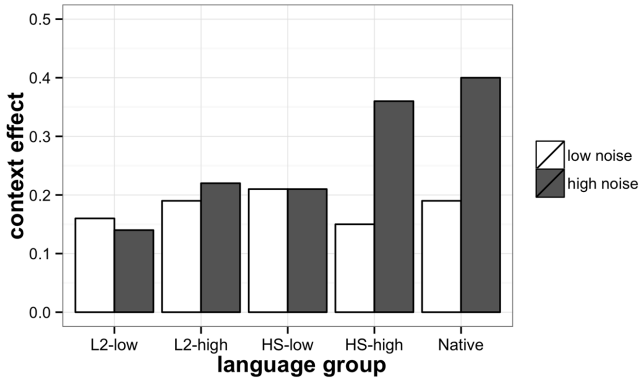


FIGURE 2 | Context effects in SPIN task for heritage, L2, and native participants. Context effect is calculated as a difference between the score in the high-predictability context condition and low-predictability context condition. 0.5 corresponds to 50% increase in accuracy in the high-predictability condition.

context condition were significant across all group comparisons ($p < 0.5$) except for between L2-low and L2-high [$t(63) = -1.58$, $p = 0.13$], L2-high and HS-low [$t(63) = 1.29$, $p = 0.2$]. In the high-noise/low context condition, the differences were significant between Native and other language groups [L2-low: $t(63) = -5.72$, $p < 0.001$, L2-high: $t(63) = -4.84$, $p < 0.001$, HS-low: $t(63) = -3.17$, $p < 0.01$, HS-high: $t(63) = -2.9$, $p < 0.01$].

To summarize, predictably, all groups benefited from low-noise compared to high-noise, however, the role of context predictability depended on the participant group and interacted with the level of noise. In the low-noise condition, there was no need in the context to recover the sentence-final word, while in the high-noise condition, the ability to efficiently process the context and to generate predictions that would help to recover the acoustically degraded sentence-final word was crucial for performance. According to the obtained results, only two groups were able to take advantage of the high-predictability context in the high-noise condition, native speakers and HSs in the high-proficiency group. These two groups relied on the context significantly more in the high-noise than in the low-noise condition. All the other groups, low-proficiency L2 and heritage, and high-proficiency L2, improved their SPIN recognition due to the high-predictability context at both noise levels to a similar, limited extent. Obviously, the high-noise/high-context condition was critical for exploring the differences among the groups, because the context was available, but the high level of noise simultaneously made the use of the context difficult. Group comparisons of accuracy scores in the critical high-noise/high-context condition reveal that native speakers are more accurate in processing SPIN than all of the other groups, and at each proficiency level, high and low, HSs outperformed L2 learners, with the L2 high-proficiency group performing similarly to the heritage low-proficiency group. Thus, HSs showed an advantage over late L2 learners, but a disadvantage compared to native speakers.

A question arises as to what deficits underlie the non-native disadvantage in late L2 learners and what aspect of SPIN processing creates advantages for HSs compared to L2 learners. In the next sections, we will briefly report the results of three experiments targeting phonological and lexical control in the same groups of participants. We will then discuss the patterns observed in the non-native populations in relation to their language learning background and setting. Two experiments tested the heritage and L2 participants' sensitivity to the phonological hardness/softness contrast that is very prominent in Russian, as it differentiates 12 pairs of Russian consonants and is widely used contrastively in building the sound shape of words and morphemes. For example, Russian infinitives and third person singular non-past tense for many verbs is contrasted by the hardness/softness of the final consonant, e.g., /pomn'it'/[4] means 'remember' while /pomn'it/ means 'he/she remembers,' with the last consonant, soft /t'/or hard /t/, providing the phonological shape for this morphosyntactically loaded contrast (Chrabaszcz and Gor, 2014). The first experiment, AXB discrimination[5], targeted lower-level perceptual sensitivity to the phonological hard/soft contrast, while the second, Picture-Word Discrimination, tested phonolexical representations, or representations of words as phonemic sequences.

EXPERIMENT 2: AXB DISCRIMINATION

MATERIAL AND METHODS

Materials

AXB discrimination test targeted the hard/soft phonological contrast in Russian consonants that has been shown to present perceptual difficulties for late American learners of Russian (Lukyanchenko and Gor, 2011; Chrabaszcz and Gor, 2014). This contrast involves the whole consonantal system of Russian with most (but not all) consonants paired according to the hard/soft feature. The hard/soft contrast is absent in English, and accordingly, English-speaking learners of Russian are not sensitive to this contrast. The test items included 84 monosyllabic consonant-vowel-consonant (CVC) non-words involving the Russian hard-soft consonant opposition (a total of 168 tokens). The stimuli were recorded by five native speakers of Russian, two males, and three females. Multi-talker speech samples ensured that the listeners would not be guided by lower-level acoustic properties of the stimuli rather than the phonological contrasts. Participants needed to process phonologically same CVC stimuli in the pronunciation of different speakers at the phonological level to establish a phonological equivalence. The study used three conditions: (1) /t-t'/ in the word-final position, as in /dot – dot'/, (2) /p-p'/ in the word-final position, as in /dop – dop'/, and (3) the /C'V-CjV/ condition, where a soft consonant in the word-initial position was contrasted with a combination of a hard consonant with a palatal /j/, as in /m'a – mja/. The contrasts and positions were selected based on the literature (Kochetov, 2002; Bondarko, 2005), and our previous research (Lukyanchenko and Gor, 2011; Chrabaszcz and Gor, 2014), as presenting the most perceptual difficulty for non-native listeners. All the available data converged on the fact that the word-final position was perceptually the most

[4]The diacritic /'/ as in /t'/is conventionally used to mark phonological softness in Russian consonants.

[5]The results of the AXB experiment were partially reported in Lukyanchenko and Gor (2011).

difficult one, and that the /t-t'/ contrast was easier than the /p-p'/ contrast. This is due to the fricativization of the soft /t'/ that provides a perceptual cue to the soft feature. There were 28 contrasts in each condition, and all contrasting consonants occurred in various vowel environments.

In order to control for the position of each token, the contrasts were grouped into triplets of four different kinds, AAB, ABB, BBA, and BAA (e.g., /mit – mit – mit'/, /mit – mit' – mit'/, etc.). The A and B items differed by the hard and soft consonants in the word-final position, and by the /C'V-CjV/ contrast in the third experimental condition. The critical token, X, always occurred in the middle. In half of the trials X corresponded to A, and in half of the trials to B. Each triplet consisted of different recordings by different speakers, and was never a repetition of the same recording by the same speaker.

Participants

The participants in AXB discrimination test were the same as in the SPIN test.

Procedure

Participants were auditorily presented with three stimuli (A, X, and B), separated by an interval of 335 ms. They were told that the first segment (A) was always different from the third segment (B), and that their task was to decide whether the second segment (X) should be categorized as A or B. Participants were required to press one of two buttons on the gamepad, left or right, to indicate whether X was identical to A or to B respectively. The next trial started 835 ms after each response. No feedback was provided. The DMDX software platform was used for stimuli presentation. The experiment took 10 min to complete.

RESULTS

A split-plot analysis of variance was used to compare the mean accuracy scores of native, heritage, and L2 speakers of Russian on the three types of contrasts: /t-t'/, /p-p'/ and /C'V-CjV/ contrast. Using an alpha level of 0.05, the results yielded a significant interaction between language group and contrast type [$F(7.91,3759.36) = 6.2$, $p < 0.01$][6], a significant within-subjects main effect of contrast type [$F(1.98,3759.36) = 24.82$, $p < 0.01$], and a significant between-subjects main effect of language group [$F(4,1899) = 91.59$, $p < 0.01$]. The effects are represented graphically in **Figure 3**. With regard to the HSs' performance, pairwise t-tests indicated that while the accuracy rate of the high-proficiency HS group was not statistically different from that of the native group [$t(63) = 1.03$, $p = 0.3$], the low-proficiency group performed significantly less accurately than the native group [$t(63) = 3.25$, $p < 0.005$]. Both L2 groups, high- and low-proficiency, were statistically less accurate than both HS groups [L2-low – HS-low: $t(63) = -7.83$, $p < 0.001$, L2-low – HS-high: $t(63) = -11.39$, $p < 0.001$, L2-high – HS-low: $t(63) = -5$, $p < 0.001$, L2-high – HS-high: $t(63) = -8.2$, $p < 0.001$] and also significantly different from each other [L2-low – L2-high: $t(63) = 2.87$, $p < 0.01$], with the low-proficiency L2 group performing the least accurately of all of the language groups.

Thus, the AXB discrimination task demonstrated the consistent advantage of proficiency-matched heritage participants compared to late L2 learners in phonological discrimination of non-word segments that had no lexical representations in the mental lexicon. Moreover, high-proficiency HSs' sensitivity to the hard/soft contrast did not differ statistically from that native speakers'. At the same time, native speakers outperformed low-proficiency HSs on all three contrasts involving hard/soft consonants.

EXPERIMENT 3: PICTURE-WORD DISCRIMINATION

The motivation behind the picture-word discrimination task was to investigate phonological processing of HSs and L2 learners of Russian in words, thereby testing the robustness of phonolexical representations. The task used minimal pairs of words, with accurate spoken word recognition depending on the discriminability of the same hard/soft contrast that was used in the AXB discrimination task. The task examined how the two populations of non-native listeners perform the mapping of the auditory input on to the stored phonological-lexical template of the word, and how their performance is similar to or different from that of native speakers of Russian.

MATERIAL AND METHODS

Materials

The stimulus materials for the task were divided into two conditions, a critical condition and a control condition. The critical condition included twelve minimal word pairs that can be distinguished based on the hardness or softness of the consonant, e.g., /mat/ 'checkmate,' versus /mat'/ 'mother.' Twelve minimal pairs were constructed for the control condition, which differed from each other in consonant voicing based on the distinction between voiced and voiceless consonants, as in /doʧ͡ka/ 'daughter,' versus /toʧ͡ka/ 'period/period,')[7]. Additionally, four distractor minimal word pairs were constructed, and two practice items were added at the beginning of the task. The materials included 30 minimal pairs (60 words). The words were mixed randomly and were recorded by a male native speaker of Russian. The words were counterbalanced between two presentation lists in such a way that both words from the same minimal pair did not occur within the same list. The same professional artist drew all 60 pictures depicting the words, so that potential differences in their visual salience would not create any biases. Lexical frequency of the words constituting the minimal pairs was not controlled, because only a few minimal pairs of nouns referring to entities that can be represented by an easily recognizable picture and differentiated by the hard/soft consonant contrast are available in Russian. However, all the words were in the high to medium frequency range.

Participants

The participants in the picture-word discrimination task were the same as in two foregoing tasks, SPIN and AXB discrimination.

[6]Mauchly's test indicated that the assumption of sphericity had been violated ($\chi = 25.7$, $p < 0.01$), therefore the degrees of freedom were corrected using Huynh–Feldt estimates of sphericity ($\varepsilon = 0.99$).

[7]The results from the control condition are not reported here.

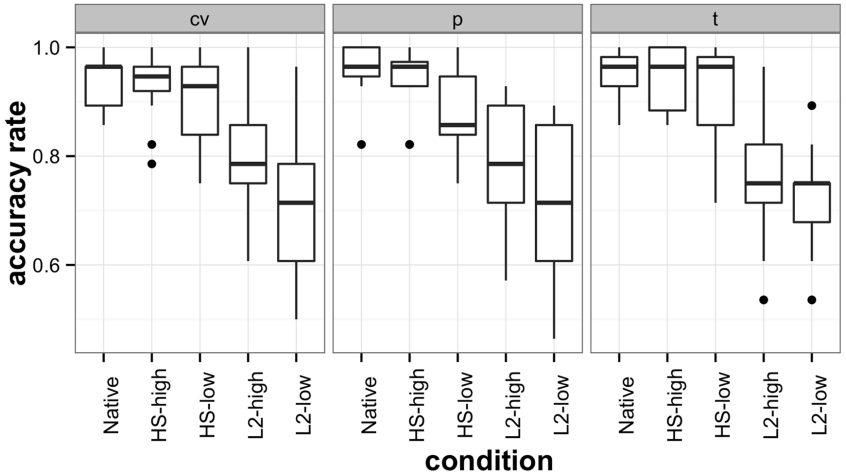


FIGURE 3 | AXB discrimination task accuracy scores across all groups. The labels 'cv', 'p', and 't' represent the three critical conditions, the C'VC-CjVC, final /p–p'/, and /t–t'/ consonants.

Procedure

Participants heard one word from the minimal pair followed by two pictures appearing on the computer screen. The presentation of the stimuli was controlled by DMDX software with the gamepad used for input. The test-takers had to decide which of the two pictures correctly matched the word that they just heard. The correct picture appeared on the left side of the screen in half of the trials, and on the right side in the other half. The participants were instructed to use the left trigger button on the gamepad to select the picture on the left, and the right trigger button to select the picture on the right. Feedback was only provided during the practice session in the beginning. The experiment took 5 min. Participants received a score of 1 selected for each test item where they selected the correct picture, otherwise they received a score of zero for the item. The accuracy scores were used for data analysis.

RESULTS

A repeated measures analysis of variance was used to compare the mean accuracy scores of native speakers ($M = 0.99$, SD $= 0.09$), high-proficiency HSs ($M = 0.98$, SD $= 0.14$), low-proficiency HSs ($M = 0.79$, SD $= 0.41$), high-proficiency L2 speakers ($M = 0.76$, SD $= 0.42$), and low-proficiency L2 speakers of Russian ($M = 0.60$, SD $= 0.49$)[8]. The results for both by-subject and by-item analyses were significant and are reported in **Table 5**.

Pairwise *t*-tests indicated that the native speakers of Russian and high-proficiency HSs of Russian were not statistically different from each other [$t(63) = 0.73$, $p = 0.47$], but were significantly different from the remaining language groups. Low-proficiency HSs performed similarly to high-proficiency L2 speakers [$t(63) = 0.38$, $p = 0.7$], and both groups were significantly more accurate in their responses than the low-proficiency L2 group [HS-low: $t(63) = 3.4$, $p < 0.05$, L2-high: $t(63) = 3.68$, $p < 0.01$, respectively]. The results are displayed graphically in **Figure 4**.

[8]One participant was excluded from the data analysis because they did not follow the task instructions; accordingly, the data analysis was done on 67 subjects.

Table 5 | Repeated measures ANOVA results for Experiment 3: picture-word discrimination.

	Df	Sum Sq	Mean Sq	*F* value	Pr(> F)
By-subject analysis					
Proficiency level	4	1.4355	0.3589	27.29	<0.000
Residuals	62	0.8153	0.0132		
By-item analysis					
Proficiency level	4	1.2704	0.3176	29.14	<0.000
Residuals	44	0.4795	0.0109		

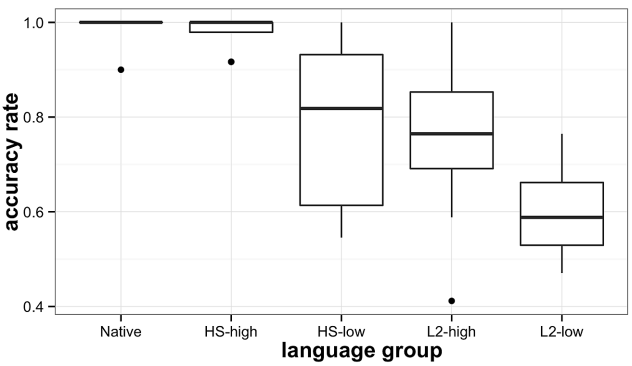


FIGURE 4 | Heritage, L2, and native participants' accuracy scores on picture-word discrimination task.

To summarize, as in the AXB discrimination task, HSs outperformed proficiency-matched late L2 learners in picture-word discrimination that involves matching the auditory input to stored phonolexical representations of words. At the same time, only the high-proficiency heritage group approximated native speakers' accuracy scores. Thus, overall, HSs show an advantage in

phonological sensitivity both in a non-word task, and in a task involving phonolexical representations of stored words. In both tasks, only the high-proficiency heritage group's accuracy scores matched native speakers' scores.

EXPERIMENT 4: LEXICAL KNOWLEDGE MULTIPLE-CHOICE TRANSLATION TEST

MATERIAL AND METHODS

When listening to sentences with a highly predictable context, listeners integrate the incoming information with the parts of the sentence that have already been processed and develop expectations about the upcoming word (Zwitserlood, 1989; Laszlo and Federmeier, 2009). The ability to develop predictions helps listeners to process the sentence-final word in high levels of noise. Given that the whole sentence is degraded due to noise, a non-native listener needs to be resistant to acoustically and phonetically degraded speech, be fast and efficient at vocabulary retrieval, and be able to quickly and efficiently generate expectations about upcoming words, integrate these expectations with what has been heard, and continue generating expectations (Mayo et al., 1997; van Wijngaarden et al., 2002; Bradlow and Alexander, 2007; Cervera and González-Alvarez, 2011). The lexical knowledge test compares vocabulary knowledge in HSs and L2 learners to determine whether the HSs' advantage in SPIN could be explained by their superior vocabulary knowledge.

Materials

The materials consisted of words in three lemma frequency ranges as determined based on Sharoff's Corpus (Russian online corpus, approximately 90 million words at the time of use[9]) that later became part of the Russian National Corpus[10]. The frequency ranges were chosen to approximate the frequencies used in a study of the M350 component, a neural response to lexical frequency (Embick et al., 2001): high-frequency, average 140 ipm (range 130–170), medium-frequency 30 ipm (30–60), and low-frequency 6 ipm (6–7). The numerical values correspond to the number of items per 1 million words of running text in the corpus (items per million, ipm). The task included nouns ($N = 30$, 15 concrete and 15 abstract), adjectives ($N = 15$), and verbs ($N = 50$), and since the target words in SPIN task were nouns, the results for this task will be briefly summarized.

Participants

The participants in the lexical knowledge multiple-choice translation test were the same as in three foregoing tasks.

Procedure

The lexical knowledge test is a multiple-choice auditory task with the Russian word presented auditorily, and three translation options in English presented visually on the computer screen. Participants were asked to choose the correct option by pressing the corresponding button on the keyboard, and their responses were recorded electronically. The experiment was run using DMDX software and took 10 min to complete.

[9]http://corpus.leeds.ac.uk/ruscorpora.html

[10]http://www.ruscorpora.ru/en/index.html

RESULTS

Based on an ANOVA, there were no statistical differences between the participant groups, heritage and L2. Conversely, the expected differences for proficiency levels and noun frequency ranges were present. Therefore, HSs showed no advantage in lexical knowledge compared to proficiency-matched late L2 learners.

DISCUSSION

The SPIN task presented the participants with Russian high- and low-predictability sentences in two levels of multi-talker babble noise, and compared their accuracy in repeating the last word of the sentence. Three groups of participants, native speakers of Russian, proficiency-matched HSs, and late L2 learners of Russian with American English as their dominant language took part in the experiment. The HSs and L2 learners were further divided into high- and low-proficiency groups based on their scores on a standardized test of oral proficiency, the oral proficiency interview, resulting in five groups. Results showed that only two groups, native speakers and high-proficiency HSs, took advantage of the high-probability context in the high-noise condition. Neither high-proficiency late L2 learners, nor low-proficiency participants improved their performance in the high-probability context. These findings must be interpreted on two levels. First, we must consider the potential role of different factors underlying the observed pattern of SPIN results across different groups. Second, the reported difference between HSs and late L2 learners needs to be connected to the language learning backgrounds in these two proficiency-matched groups.

Given that the study used multi-talker babble noise that combined energetic and informational masking, one can expect an impact of noise on all levels of speech processing, from low-level acoustic-phonetic interference to high-level sentence integration. Which levels were responsible for the observed differences in heritage and L2 processing of SPIN, and what features/aspects of their respective learning backgrounds could have contributed to the differences in processing degraded speech? Two experiments targeting phonological processing and one experiment testing lexical knowledge with the same participant groups as the ones that took part in the SPIN test make it possible to gage phonological and lexical knowledge of the heritage and late L2 learners and compare the outcomes to the results of the SPIN test. AXB discrimination used non-word CVC syllables and assessed participants' sensitivity to the hard/soft phonological contrast in Russian consonants. The picture-word discrimination task tested the same contrast in minimal pairs of words and examined the robustness of phonolexical representations differentiated by this contrast. Both tasks produced the same results: HSs outperformed proficiency-matched L2 learners, and high-proficiency HSs' accuracy scores were statistically similar to those of native speakers. HSs showed both a phonological and phonolexical advantage over L2 learners, and the high-proficiency heritage group approximated phonological sensitivity and the robustness of phonolexical representations of native speakers. Importantly, no direct statistical comparisons were made due to the differences in the design and the variables included in the three experiments. Therefore, one could argue that the comparisons between

the outcomes of SPIN and the phonological tasks are suggestive rather than conclusive. The logic behind the comparisons of the global outcomes of the experiments was that both phonological tasks targeted the most difficult Russian phonological contrast for American learners, and crucially, the most pervasive one in the Russian consonantal system (Chrabaszcz and Gor, 2014). They gaged non-native sensitivity to different positional and contextual allophones, in words and non-words, and thereby provided a global assessment of phonological control. Conversely, no advantage of HSs over L2 learners was demonstrated in the lexical knowledge test, and therefore their superior lexical knowledge and/or better lexical entrenchment should be discarded as a possible explanation of the heritage advantage in SPIN.

Note that if this phonological advantage were the only cause of the differences in the SPIN results, there would have been no interaction between the level of noise and context predictability in the heritage or any other group. Conversely, two groups show a noise/context interaction, and these are native speakers and high-proficiency HSs. At the same time, the high-proficiency L2 group approximates the accuracy scores of the heritage low-proficiency group and does not show any noise/context interaction in SPIN. In other words, high-proficiency helps HSs to take advantage of the high-predictability context in the high-noise condition (as do native speakers), but the context does not help proficiency-matched late L2 learners. It is exactly this context/noise interaction with heavier reliance on the context only in high-proficiency HSs and native speakers of Russian that indicate that, indeed, superior phonological decoding abilities combined with the efficiency in quickly integrating the incoming information with the preceding sentence context and generating predictions about the upcoming word are the properties of native and high-proficiency heritage processing of speech in adverse conditions. Importantly, while HSs' advantage over L2 learners has been documented both in SPIN and the tests of phonological sensitivity, only high-proficiency heritage listeners approximate native speech recognition in adverse conditions. In order to do so, they should be able to generate predictions by relying on efficient bottom–up and top–down mechanisms of speech processing, phonological decoding, and sentence integration that act in a bootstrapping fashion.

Heritage speakers acquire their language as native speakers, in a naturalistic environment, since birth, from their caregivers, and as their first language. This language learning background should provide them with a native advantage in auditory speech recognition demonstrated in studies of native and non-native speech perception (see Introduction). At the same time, HSs differ from native speakers and balanced bilinguals in that their proficiency in what was chronologically their first language is non-native. This is why they can be matched in proficiency with late L2 learners. If an early naturalistic start from birth always provides an advantage to HSs over late L2 learners, regardless of the proficiency level, both heritage groups should outperform both L2 groups. If language proficiency also matters, first, high-proficiency HSs will outperform low-proficiency ones, and second, there could be a proficiency-based effect that will be observed only in one proficiency range, but not the other.

The outcomes of the SPIN experiment supported both the early advantage and the role of proficiency in HSs. Thus, HSs outperformed late L2 learners, and his advantage was present both in the high-proficiency and low-proficiency-based comparisons. The results for the critical high-noise and high context predictability condition, where the differences among the groups were the most salient, show an advantage of HSs over L2 learners, thereby supporting the early heritage advantage. At the same time, the study found that the ability to profit from the high-predictability context in the high-noise condition was mitigated by proficiency in HSs. Only the high-proficiency heritage group showed native-like reliance on context probabilities in high-noise. This latter finding speaks to the role of proficiency in heritage as well as L2 listeners, given that proficiency-matched late L2 learners consistently lag behind HSs.

The reasons why HSs' proficiency in their heritage language falls short of native proficiency are beyond the scope of this study and are still widely debated. These are predominantly the developmental factors, attrition, incomplete acquisition or, frequently, a combination of both, with their relative weight depending on the age of reduced exposure to the heritage language, and the amount of exposure to L1 and L2 since the age of L2 onset (Montrul, 2008, 2012; Bylund, 2009; Keijzer, 2010; Schmid, 2010; Polinsky, 2011). The age of L2 onset, while a decisive factor, is not the only one; the relative amount of time each of the two languages is used by a HS is no less important. Crucially, language aptitude is positively correlated with such aspects of heritage language proficiency as grammatical knowledge, as assessed by a grammaticality judgment test (Bylund et al., 2010). The study by Bylund and colleagues targeted prepubescent attriters who experienced a break with their Spanish L1 (heritage language) before puberty and switched to Swedish. In the participants with below-average aptitude, scores on the grammaticality judgment test positively correlated with the amount of daily use of Spanish. Heritage language proficiency level is thus a product of a complex interplay of cognitive, social and environmental factors. It is influenced by the amount of exposure to, and the level of engagement with, each of the two languages, which interact with language aptitude.

It should be noted that establishing language proficiency is both a theoretical and a practical problem, and the current study uses the standard of global language proficiency testing, an oral proficiency interview. It establishes the level based on the ability of the participant to perform language functions, such as being able to narrate an event in major time frames or handle a situation with a complication in a role-play situation. This format is suitable for testing HSs (see Kagan and Friedman, 2003; Polinsky and Kagan, 2007), it avoids the bias of visually presented reading materials, and since the oral proficiency interview has an interactive format, it tests both receptive (listening) and productive (speaking) skills. The assessment of the global ability to speak and interact with a conversational partner in order to establish a level of language proficiency is ecologically valid, given that language is primarily a means of communication. At the same time, the results of the study raise the issue of an asymmetry between the levels of performance on separate linguistic aspects, such as phonological sensitivity or speech recognition in adverse conditions and the

global proficiency rating established in a conversational format. According to the obtained results, groups of non-native speakers with different learning profiles, but matched on speaking proficiency may still differ in their control of individual aspects of linguistic performance, a finding that needs further research. This fairly plausible finding documents the specific role of early naturalistic language exposure in shaping the mechanisms underlying phonological processing and speech recognition.

The results of the study shed new light on the role of early naturalistic experience in learning a heritage language, the first language to be learned chronologically, but the second in dominance, and characterized by non-native levels of proficiency. The reported results suggest that early naturalistic language learning experience is necessary, but insufficient for developing native listening strategies that ensure robust speech recognition in adverse conditions. Listeners encounter different forms of degraded speech in their everyday experience, ranging from band-pass filtered speech in phone communications to listening to speech in noisy conditions and separating the speech stream of the interlocutor from that of another individual speaking at the same time. It appears that a high-proficiency in the heritage language is necessary for robust speech recognition in adverse conditions. The length and intensity of exposure to, and use of the heritage language initially acquired naturalistically most likely mitigate the high or low heritage language proficiency level. To the best of our knowledge, this is the first study devoted to speech recognition under adverse conditions in HSs. It has established that high-proficiency HSs outperform oral proficiency-matched late L2 learners on the use of contextual information for disambiguation of sentence-final words in sentences presented in multi-talker babble noise. The group of high-proficiency HSs was not statistically different from native speakers in their use of contextual information.

While the project was not designed as a correlational study, the outcomes of the SPIN test can be compared with the results of two tests of phonological control and a test of lexical knowledge completed by the same groups of participants. These *post hoc* comparisons are justified given the previous findings regarding the role of L2 phonology and lexical-semantic knowledge in SPIN processing. At the same time, since no direct statistical comparisons were made across the test results, the outcomes are interpreted within the context of what is known about the factors contributing to L2 deficits in SPIN. Reduced L2 phonological sensitivity has an established record of being associated with non-native problems with SPIN processing in non-word sequences (Garcia Lecumberri and Cooke, 2006; Cutler et al., 2008; Broersma and Scharenborg, 2010) and words (Oyama, 1982; Meador et al., 2000; Cooke et al., 2008). The role of lexical-semantic knowledge and a semantic context advantage for native, but not non-native speakers processing SPIN has also been demonstrated (Golestani et al., 2009, 2013; Hervais-Adelman et al., 2014). A comparison of the results of the SPIN test with two tests of phonological control, a low-level phonemic AXB discrimination task and a picture-word discrimination task showed a clear advantage of HSs over L2 learners in all three tasks. Both AXB and the picture-word discrimination task, which tested the robustness of phonolexical representations of minimal pairs of words, targeted the same phonological contrast, the hard-soft consonant distinction that is both difficult for American learners of Russian and very pervasive in Russian speech. Therefore, the performance on this contrast can be considered as a measure of L2 phonological control. Conversely, a multiple choice test of lexical knowledge did not show a heritage advantage over L2 learners. Importantly, in all three tasks, SPIN and two tests of phonological sensitivity, only high-proficiency HSs approximated native performance. The study asks the question whether the SPIN advantage can be explained away by the phonological advantage in HSs. It concludes that phonological sensitivity contributes to the heritage advantage in SPIN, but that more than just phonological sensitivity underlies the heritage advantage. This conclusion is based on the observation that HSs outperform L2 learners on all three tasks in both proficiency ranges, but only high-proficiency HSs are able to use the contextual information to disambiguate the sentence-final word in noise. The study argues that phonological sensitivity and high-level processing involving sentence integration and prediction generation act in a bootstrapping fashion. This leads to a qualitatively different nativelike use of sentence context in noise in the high-proficiency heritage group.

However, at this stage, this conclusion remains tentative, and it invites further research. A promising direction is to continue the work of Aydelott and Bates (2004) by exploring both facilitation and inhibition in non-native processing of speech in adverse conditions, and manipulating the type of masking, low-level energetic versus high-level informational. This will ultimately make it possible to arrive at a better-informed conclusion about the role of top–down processes of sentence integration and generation of predictions about the upcoming word in early and late learners. Another direction in the behavioral and neurolinguistic study of non-native populations, including HSs, is the use of retroactive auditory semantic priming experiments, when the prime word is presented in noise, and the target is either semantically related or unrelated to the prime, e.g., 'parrot – bird' is a related pair, and 'parrot – cake' is an unrelated one (Golestani et al., 2009; Hervais-Adelman et al., 2014). The participant hears the prime presented in noise, and the target in clear conditions, and must decide which of the two visually presented words was the prime. The facilitative role of the retroactive semantic context in native, but not non-native language was observed in reverse semantic priming experiments for native French speakers listening to French and English word pairs (Golestani et al., 2009). Additionally, only a native language context effect was found in an fMRI study using the same retroactive semantic priming technique (Hervais-Adelman et al., 2014). The use of retroactive semantic priming in noise makes it possible to tease apart prediction from integration of context information and to test non-native use of both in degraded speech recognition. Both these directions have a potential to lead to new insights with regard to heritage and late L2 speech processing.

ACKNOWLEDGMENTS

This study was funded by the Center for Advanced Study of Language at the University of Maryland. I express my deep gratitude to the people who contributed to this project by providing their input, and assistance with programming the experiments, data acquisition, and analyses: Anna Chrabaszcz, Svetlana Cook, Scott Jackson, and Michael Long.

REFERENCES

Abrahamsson, N., and Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: listener perception versus linguistic scrutiny. *Lang. Learn.* 59, 249–306. doi: 10.1111/j.1467-9922.2009.00507.x

Aydelott, J., and Bates, E. (2004). Effects of acoustic distortion and semantic context on lexical access. *Lang. Cogn. Process.* 19, 29–56. doi: 10.1080/01690960344000099

Bhatia, T. K., and Ritchie, W. C. (2013). *The Handbook of Bilingualism and Multilingualism*. Oxford: Wiley-Blackwell.

Bondarko, L. (2005). Phonetic and phonological aspects of the opposition of "soft" and "hard" consonants in the modern Russian language. *Speech Commun.* 47, 7–14. doi: 10.1016/j.specom.2005.03.012

Bradlow, A. R., and Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J. Acoust. Soc. Am.* 121, 2339–2349. doi: 10.1121/1.2642103

Bradlow, A. R., and Bent, T. (2002). The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.* 112, 272–284. doi: 10.1121/1.1487837

Broersma, M., and Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise. *Speech Commun.* 52, 980–995. doi: 10.1016/j.specom.2010.08.010

Bylund, E. (2009). Maturational constraints and first language attrition. *Lang. Learn.* 59, 687–715. doi: 10.1111/j.1467-9922.2009.00521.x

Bylund, E., Abrahamsson, N., and Hyltenstam, K. (2010). The role of language aptitude in first language attrition: the case of prepubescent attriters. *Appl. Ling.* 31, 443–464. doi: 10.1093/applin/amp059

Bylund, E., Abrahamsson, N., and Hyltenstam, K. (2012). Does L1 maintenance hamper L2 nativelikeness? A study of L2 ultimate attainment in early bilinguals. *Stud. Sec. Lang. Acquis.* 34, 215–241. doi: 10.1017/S0272263112000034

Cervera, T., and González-Alvarez, J. (2011). Test of Spanish sentences to measure speech intelligibility in noise conditions. *Behav. Res.* 43, 459–467. doi: 10.3758/s13428-011-0063-2

Chang, C. B., Yao, Y., Haynes, E. F., and Rhodes, R. (2011). Production of phonetic and phonological contrast by heritage speakers of Mandarin. *J. Acoust. Soc. Am.* 129, 3964–3980. doi: 10.1121/1.3569736

Chrabaszcz, A., and Gor, K. (2014). Context effects in the processing of phonolexical ambiguity in L2. *Lang. Learn.* 64, 415–455. doi: 10.1111/lang.12063

Cook, V. (2003). *The Effects of the Second Language on the First*. Clevedon: Multilingual Matters.

Cooke, M., Garcia Lecumberri, M. L., and Barker, J. (2008). The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. *J. Acoust. Soc. Am.* 123, 414–427. doi: 10.1121/1.2804952

Cutler, A., Garcia Lecumberri, M. L., and Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: effects of local context. *J. Acoust. Soc. Am.* 124, 1264–1268. doi: 10.1121/1.2946707

Cutler, A., Weber, A., Smits, R., and Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *J. Acoust. Soc. Am.* 116, 3668–3678. doi: 10.1121/1.1810292

DeKeyser, R. (2013). Age effects in second language learning: stepping stones toward better understanding. *Lang. Learn.* 63, 52–67. doi: 10.1111/j.1467-9922.2012.00737.x

Embick, D., Hackl, M., Schaeffer, J., Kelepir, M., and Marantz, A. (2001). A magnetoencephalographic component whose latency reflects lexical frequency. *Cogn. Brain Res.* 10, 345–348. doi: 10.1016/S0926-6410(00)00053-7

Forster, K. I., and Forster, J. C. (2003). DMDX: a Windows display program with millisecond accuracy. *Behav. Res. Methods Instr. Comp.* 35, 116–124. doi: 10.3758/BF03195503

Garcia Lecumberri, M. L., and Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am.* 119, 2445–2454. doi: 10.1121/1.2180210

Golestani, N., Hervais-Adelman, A., Obleser, J., and Scott, S. K. (2013). Semantic versus perceptual interactions in neural processing of speech-in-noise. *Neuroimage* 79, 52–61. doi: 10.1016/j.neuroimage.2013.04.049

Golestani, N., Rosen, S., and Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: the contribution of semantics. *Biling. (Camb. Engl.)* 12, 385–392. doi: 10.1017/S1366728909990150

Gor, K. (2000). "Experimental research of vowel reduction in Russian: implications for inter language phonology and for teaching Russian pronunciation," in *The Learning and Teaching of Slavic Languages and Cultures: Toward the 21st Century*, eds O. Kagan and B. Rifkin (Columbus, OH: Slavica), 193–214.

Gor, K., and Cook, S. (2010). Non-native processing of verbal morphology: in search of regularity. *Lang. Learn.* 60, 88–126. doi: 10.1111/j.1467-9922.2009.00552.x

Gor, K., Cook, S., Malyushenkova, V., and Vdovina, T. (2009). Verbs of motion in highly proficient learners and heritage speakers of Russian. *Slav. East Eur. J.* 53, 386–408.

Hervais-Adelman, A., Pefkou, M., and Golestani, N. (2014). Bilingual speech-in-noise: neural bases of semantic context use in the native language. *Brain Lang.* 132, 1–6. doi: 10.1016/j.bandl.2014.01.009

Kagan, O., and Friedman, D. (2003). Using the OPI to place heritage speakers of Russian. *Foreign Lang. Ann.* 36, 536–545. doi: 10.1111/j.1944-9720.2003.tb02143.x

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337–1351. doi: 10.1121/1.381436

Keating, G., Jegerski, J., and VanPatten, B. (2011). Who was walking on the beach? Anaphora resolution in Spanish heritage speakers and adult second language learners. *Stud. Sec. Lang. Acquis.* 33, 193–221. doi: 10.1017/S0272263110000732

Keijzer, M. (2010). The regression hypothesis as a framework for first language attrition. *Biling. Lang. Cogn.* 13, 9–18. doi: 10.1017/S1366728909990356

Kochetov, A. (2002). *Production, Perception, and Emergent Phonotactic Patterns: A Case of Contrastive Palatalization*. New York: Routledge.

Laszlo, S., and Federmeier, K. (2009). A beautiful day in the neighborhood: an event-related potential study of lexical relationships and prediction in context. *J. Mem. Lang.* 61, 326–338. doi: 10.1016/j.jml.2009.06.004

Lee-Ellis, S.-Y. (2011). The elicited production of Korean relative clauses by heritage speakers. *Stud. Sec. Lang. Acquis.* 33, 57–89. doi: 10.1017/S0272263110000537

Lekic, M. D., Davidson, D. E., and Gor, K. S. (2008). *Russian Stage One: Live from Russia!* Dubuque, IA: Kendall/Hunt Publishing Company.

Long, M. H., Gor, K., and Jackson, S. (2012). Linguistic correlates of second language proficiency: proof of concept with ILR 2-3 in Russian. *Stud. Sec. Lang. Acquis.* 34, 99–126. doi: 10.1017/S0272263111000519

Lubensky, S., Ervin, G., McLellan, L., and Jarvis, D. (2002). *Nachalo*. New York, NY: McGraw-Hill.

Lukyanchenko, A., and Gor, K. (2011). Perceptual correlates of phonological representations in heritage speakers and L2 learners, in *Proceedings of the Thirty Fifth Annual Boston University Conference on Language Development* (Somerville, MA: Cascadilla Press), 414–426.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). Speech recognition in adverse conditions: a review. *Lang. Cogn. Process.* 27, 953–978. doi: 10.1080/01690965.2012.705006

Mayo, L. H., Florentine, M., and Buus, S. (1997). Age of second language acquisition and perception of speech in noise. *J. Speech Lang. Hear. Res.* 40, 686–693. doi: 10.1044/jslhr.4003.686

Meador, D., Flege, J. E., and Mackay, I. R. A. (2000). Factors affecting the recognition of words in a second language. *Biling. Lang. Cogn.* 3, 55–67. doi: 10.1017/S1366728900000134

Montrul, S. (2008). *Incomplete Acquisition in Bilingualism. Re-examining the Age Factor*. Amsterdam: John Benjamins. doi: 10.1075/sibil.39

Montrul, S. (2009). Incomplete acquisition of tense-aspect and mood in Spanish heritage speakers. *Int. J. Biling.* 13, 239–269. doi: 10.1177/1367006909339816

Montrul, S. (2011). Morphological errors in Spanish second language learners and heritage speakers. *Stud. Sec. Lang. Acquis.* 33, 163–192. doi: 10.1017/S0272263110000720

Montrul, S. (2012). Is the heritage language like a second language? *EUROSLA Yearb.* 12, 1–29. doi: 10.1075/eurosla.12.03mon

Montrul, S., Davidson, J., de la Fuente, I., and Foote, R. (2013). The role of experience in the acquisition and production of diminutives and gender in Spanish: evidence from L2 learners and heritage speakers. *Sec. Lang. Res.* 29, 87–118. doi: 10.1177/0267658312458268

Montrul, S., de la Fuente, I., and Davidson, J., and Foote, R. (2014). Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. *Biling. Lang. Cogn.* 17, 118–138. doi: 10.1017/S1366728913000114

Montrul, S., Foote, R., and Perpiñan, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: the effects of age and context of acquisition. *Lang. Learn.* 58, 3–54. doi: 10.1111/j.1467-9922.2008.00449.x

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Stud. Sec. Lang. Acquis.* 20, 139–154. doi: 10.1017/S0272263198002022

Oh, J., Jun, S., Knightly, L., and Au, T. (2003). Holding on to childhood language memory. *Cognition* 86, B53–B64. doi: 10.1016/S0010-0277(02)00175-0

Oliver, G., Gullberg, M., Hellwig, F., Mitterer, H., and Indefrey, P. (2012). Acquiring L2 sentence comprehension: a longitudinal study of word monitoring in noise. *Biling. Lang. Cogn.* 15, 841–857. doi: 10.1017/S1366728912000089

Oyama, S. (1982). "A sensitive period for the acquisition of a nonnative phonological system," in *Child-Adult Differences in Second Language Acquisition*, eds S. Krashen, R. Scarcella, and M. Long (Rowley, MA: Newbury House), 20–38.

Pinet, M., Iverson, P., and Huckvale, M. (2011). Second-language experience and speech-in-noise recognition: effects of talker-listener accent similarity. *J. Acoust. Soc. Am.* 130, 1653–1662. doi: 10.1121/1.3613698

Polinsky, M. (2008). "Heritage language narratives," in *Heritage Language Education: A New Field Emerging*, eds D. M. Brinton, O. Kagan, and S. Bauckus (New York: Routledge), 149–164.

Polinsky, M. (2011). Reanalysis in adult heritage language: new evidence in support of attrition. *Stud. Sec. Lang. Acquis.* 33, 305–328. doi: 10.1017/S027226311000077X

Polinsky, M. (2014). When L1 becomes an L3: Do heritage speakers make better L3 learners? *Biling. Lang. Cogn.* 1–16. doi: 10.1017/S1366728914000315

Polinsky, M., and Kagan, O. (2007). Heritage languages: in the 'wild' and in the classroom. *Lang. Ling. Comp.* 1, 368–395. doi: 10.1111/j.1749-818X.2007.00022.x

Psaltou-Joyceya, A., and Kantaridoub, Z. (2011). Major, minor, and negative learning style preferences of university students. *System* 39, 103–112. doi: 10.1016/j.system.2011.01.008

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robin, R., Evans-Romaine, K., and Shatalina, G. (2014). *Golosa: A Basic Course in Russian*. Upper Saddle River, NJ: Prentice Hall.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Appl. Psycholing.* 27, 465–485. doi: 10.1017/S014271640606036X

Rogers, T. C. (2014). *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Schmid, M. (2010). Languages at play: the relevance of L1 attrition to the study of bilingualism. *Biling. Lang. Cogn.* 13, 1–7. doi: 10.1017/S1366728909990368

Shtern, A. S. (1992). *Perceptual Aspects of Speech Processing*. Saint Petersburg: Saint Petersburg University Press.

Shtern, A. S. (2001). *Russian Articulatory Tables*. Saint Petersburg: Saint Petersburg University Press.

Thompson, I. (2000). "Assessing foreign language skills: data from Russian," in *The Learning and Teaching of Slavic Languages and Cultures*, eds O. Kagan and B. Rifkin (Bloomington, IN: Slavica), 255–284.

Valdés, G. (2005). Bilingualism, heritage language learners, and SLA research: opportunities lost or seized? *Mod. Lang. J.* 89, 410–426. doi: 10.1111/j.1540-4781.2005.00314.x

Van Engen, K. J., and Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *J. Acoust. Soc. Am.* 121, 519–526. doi: 10.1121/1.2400666

van Wijngaarden, S. J., Steeneken, H. J., and Houtgast, T. (2002).Quantifying the intelligibility of speech in noise for nonnative talkers. *J. Acoust. Soc. Am.* 112, 3004–3013. doi: 10.1121/1.1456928

Zwitserlood, P. (1989). The locus of the effects of the sentential-semantic context in spoken word processing. *Cognition* 32, 25–64. doi: 10.1016/0010-0277(89)90013-9

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 July 2014; accepted: 26 November 2014; published online: 19 December 2014.

*Citation: Gor K (2014) Raspberry, not a car: context predictability and a phonological advantage in early and late learners' processing of speech in noise. Front. Psychol. **5**:1449. doi: 10.3389/fpsyg.2014.01449*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 11 July 2014
doi: 10.3389/fpsyg.2014.00716

In search of conceptual frameworks for relating brain activity to language function

Mike A. Sharwood Smith*

Moray House School of Education, University of Edinburgh, Edinburgh, UK
**Correspondence: msharwood@blueyonder.co.uk*

Edited by:
Christos Pliatsikas, University of Kent, UK

Reviewed by:
Laura Sabourin, University of Ottawa, Canada

Keywords: multilingual, bilingualism, second language acquisition, second language processing, neurolinguistics, psycholinguistics, models, theoretical

The focus of the current topic is the analysis and interpretation of second language (L2) and multilingual data. Looking at data from speakers who have learned their additional languages after the mother tongue has become well established is of special interest. It advances our knowledge about how different language systems share space in the same mind, a question to be asked of any kind of multilingual at any age—and secondly it can tell us more about potential differences between early and later learned languages (Kim et al., 1997; Kovelman et al., 2008). More recent research points to brain areas activated in late learners of L2s becoming more and more like those of L1 acquirers as their proficiency advances (Green, 2003). At earlier stages of acquisition, adults may simply adopt compensatory strategies, for example recruiting new cognitive resources that have become available with increasing maturity to complete communicative tasks that are demanding either because, unlike very young children, they personally want or are compelled by interlocutors to communicate complex ideas, or because the requirements of a given experiment simply make the tasks demanding. This will therefore implicate regions of the brain that are much less involved in young language users and these may stay involved even where higher levels of L2 ability render them much less important or even unnecessary. In any case, to get the full picture we need ways of tracking this strategic activity, one aspect of which is the deployment of explicit processes, both those involving conscious awareness and those that may be raised to awareness but can also operate subconsciously but there will surely be processes that only operate subconsciously as well (Sharwood Smith and Truscott, 2011). These will affect not only the spontaneous uses of L1 and L2 but also performance on experimental tasks. Tracking brain activity with sophisticated apparatus is not enough of course: the data needs to be analyzed and for this we need very sophisticated theoretical frameworks to guide interpretation.

While research techniques such as brain imaging are gradually acquiring greater precision, helping to reveal much more about brain activity associated with linguistic processing, there still remain many problems interpreting results. This may not be an immediate problem in a given experiment because the research question may be suitably precise and focussed enough to guarantee an answer of sorts in the hope that answers to limited questions may gradually accumulate and provide the basis for wider explanations. In this way, for example, syntactic and semantic processing can be teased apart on the basis of participants' differing responses to examples of, respectively, syntactic and semantic anomaly which then allows researchers to identify separate neural signatures and provide support for particular accounts of the status of language vis-a-vis other types of cognition. Issues of interpretation become more evident when trying to put results into a wider explanatory context. One problem concerns the choice of which theory and which concepts and categories to import from a neighboring research domain. Another one, related to that, is locating conceptual models and frameworks in related domains (neurolinguistic, psycholinguistic, theoretical linguistic) that can be combined in such a way as to promise the best possible explanation.

Assuming the focus is on explaining language, and leaving aside sociolinguistic issues, if we get down to the basics, what do we have? 1.3 kilos (three pounds) of soft tissue and our current understanding of its functional architecture. Add to that theories about psychological function and, in many cases at least, entirely separate, well developed theories about linguistic structure. Each of these theories has, for very good reasons, its own conceptual framework and terminology, and its own favorite methods of investigation. For satisfactory explanations of how the brain stores and processes languages, we need somehow to coordinate findings in all these different disciplines. At the same time, it is not a straightforward job to bootstrap, for example, a Minimalist approach to explaining language structure to a model of human memory and make it into a real-time processing theory. This is true notwithstanding the obvious need, in the elaboration of theories of processing and development, for fine-grained accounts of linguistic structure. Standard generative linguistic approaches to language employ terms and concepts to explain abstract linguistic structure that are outside time and space. Without going into the details, these are notions like "move," a structure-building operation changing the position of some item in a structure, "merge," a combining operation, and "feature-checking," the process whereby two associated items in a structure are assessed to see if they can be "licensed," i.e., co-exist in their current position (in

a particular manner determined by the theory). The use of such spatiotemporal metaphors, without any responsibility for explaining processing facts, is a highly effective device for describing structural relationships in syntax. At the same time these metaphors should not be translated straightforwardly into real-time, on-line processing terms. If they were thus interpreted, that would constitute a serious category error—confusing abstract theoretical linguistic concepts for ones that are custom designed to describe and explain events in real time. Either that, or it would constitute a new and different type of claim entirely, i.e., that the abstract concepts can do double duty and describe real-time events *as well*. Unless there is such a claim, working with linguistic theory means operating at a different level of description from *psycho*linguistic processing theory. Yet another level, distinct from both of those ones, seeks to provide *neuro*linguistic descriptions and explanations (Sabourin and Haverkort, 2003). Again, there is no guaranteed simple and straightforward translation of psychological explanations into neurofunctional ones either. In a psychological description, a working memory (WM) may be described as, say, a single module where its neural substrate is seen as being distributed over, say, three different systems, each located in different areas of the brain. If the model of memory is a modular one, the number of subsystems can still be different depending on whether the description is psychological or neurological.

Although functional models in psycholinguistics do not have to match their linguistic and neurolinguistic counterparts in any literal way, the need to interpret one into terms of the other does have to be acknowledged so that the search for, or development of compatible models across disciplines can proceed. While research continues sorting out the "easier" problems of data collection and analysis within each of these three disciplines, it is still useful to cast a critical eye on many of the basic assumptions being made and raise some questions about them. It is fair to point out that with the increasing sophistication of techniques that record brain activity, the problem of resolving the bigger questions will gradually become more tractable but only provided that suitable theoretical, compatible, well-founded models in companion domains can be identified so that the bigger questions can be formulated.

To take memory as a case in point, what is the relationship between on-line language processing, in this case by bilinguals (multilinguals) and the formation of new stable memories? How and where are the relevant memories formed? Should different types of memory be assumed as is often the case nowadays. If so, how many? Let us begin, say, with the initial registering of the acoustic stream: if we can accept that auditory-acoustic memories are formed in the primary auditory cortex, where and how is the subset of those memories that are identified as language-relevant processed further, thereby forming (some claim) separate types of memory? Is it legitimate to talk, for example, of a single separate "linguistic memory" system or are there in fact two separate types of memory involved, phonological and syntactic (Jackendoff, 2002; Truscott and Sharwood Smith, 2004; Sharwood Smith and Truscott, 2014)? Moving on to WM, is this part of a unified system serving all types of cognitive and perceptual activity or is WM also modular and domain-specific? If so, which modules and which domains are we talking about? And, during repeated on-line processing, when items that have appeared in (one or other instances of) WM have eventually become stable and established items in longer term memory (LTM), should we treat this acquisition process as resulting from the successful transition of the relevant items from one memory system (WM) into another one (LTM) or, alternatively, should we treat WM and LTM as part of a single memory system thereby characterizing acquisition as establishment of an enduring trace in LTM that then becomes increasingly accessible over time (Cowan, 2005; Baddeley, 2012)? It might not matter which option you choose for some purposes but if the models are going to be useful they may each have different empirical consequences when applied to the more complex questions of linguistic acquisition and performance. The plain fact is that models being used today still do not yet specify exactly how language systems are stored and used within one mind/brain. A much more detailed architecture is required to required to meet this requirement and explain how discourse/pragmatic, semantic, morphosyntactic and semantic features are stored and interact across a single or across multiple language systems (see, for example, proposals in Sharwood Smith and Truscott, 2014).

Connected with the decisions about which model of memory and storage to use is the question whether or not there is anything like a "language acquisition device" and if there is one, how does it work? What is its neural substrate? To take representational models of cognition, for example, some assign a special status to human language while others treat it as part of general cognition. The emergentist architecture proposed by O'Grady is an example of the latter (O'Grady, 2000). O'Grady explains language acquisition as cognitive development that is driven by the selection of ever more efficient processing operations to handle the input. One such operation seeks to minimize the burden on WM. Sharwood Smith and Truscott's account is similar at least in this one respect, denying the need for a language acquisition device, whereas Carroll's Autonomous Induction Model is different and posits a modular, failure-driven acquisition mechanism that is unique to language (Carroll, 2001, 2007; see discussion in Truscott and Sharwood Smith, 2004; Sharwood Smith et al., 2012; Sharwood Smith and Truscott, 2014). In any study of acquisition, it is fair to ask what background theoretical commitments the researchers are making and to what extent it is a matter of principled choice or just one of convenience, understandable as that might be.

My final example is the notion "representation." If we set aside non-symbolic accounts, somewhere along the line we have to have a clear idea of how to treat representations at the different levels of description (and explanation) that we have been dealing with, ranging from simple ones like "word," "syllable," and "lexical item" to ones like "noun," phonological and syntactic "features" and the whole gamut of theoretical categories that we wish to deploy in some form or other for experimental investigation and data analysis. Representations may be psychological constructs but they should have neural

correlates. For example, Damasio's notion of "dispositional representation" meshes easily with the way linguistic or psychological representations are conceived. A dispositional representation is "a potential pattern of neuron activity in small ensembles of neurons and may be distributed over a number of different locations in the cortex, the precise locations depending on the type of representation and whether it is innate or acquired as a result of experience" (Damasio, 1994, pp. 102–105). This provides another illustration of how the neural equivalent of a psychological representation located in one particular place in a theoretical model can be a structure that is distributed across the neural system in different places. It also shows, incidentally, that you do not need to choose between symbolic representational accounts on the one hand and connectionist accounts based on (biological) neural networks on the other. Networks, representations and modular architectures can live peacefully together.

To some extent this short discussion is more a look into the future than a critique of past and present research. It is somewhat of a cliché to say research in this area needs to be conducted by teams from different research domains. To some extent this is already happening. My basic point is that the development of useful conceptual frameworks that can support such multidisciplinary research is still in its infancy. I have my own suggestions about what such a framework might look like but that is another story.

ACKNOWLEDGMENT

My thanks to an anonymous reviewer and John Truscott for their comments on previous versions of this manuscript.

REFERENCES

Baddeley, A. (2012). Working memory: theories, models, and controversy. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Carroll, S. (2001). *Input and Evidence: The Raw Material of Second Language Acquisition*. Amsterdam: Benjamins.

Carroll, S. (2007). "Autonomous induction theory," in *Theories in Second Language Acquisition: An Introduction*, eds B. VanPatten and J. Williams (New York, NY: Routledge), 155–174.

Cowan, N. (2005). *Working Memory Capacity*. New York, NY: Psychology Press. doi: 10.4324/9780203342398

Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Papermac.

Green, D. (2003). "Neural basis of lexicon and grammar in L2 acquisition: the convergence hypothesis," in *The Lexicon–Syntax Interface in Second Language Acquisition*, eds R. van Hout, A. Hulk, F. Kuiken, and R. Towell (Amsterdam: John Benjamins), 197–218.

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

Kim, K., Relkin, N., Lee, K.-M., and Hirsch, J. (1997). Distinct cortical areas associated with native and second languages. *Nature* 3, 171–174.

Kovelman, I., Baker, S., and Petitto, L.-A. (2008). Bilingual and monolingual brains compared: a functional magnetic resonance imaging investigation of syntactic processing and a possible "neural signature" of bilingualism. *J. Cogn. Neurosci.* 20, 153–169. doi: 10.1162/jocn.2008.20011

O'Grady, W. (2000). *Syntactic Carpentry: An Emergentist Approach to Syntax*. Mahwah, NJ: Erlbaum.

Sabourin, L., and Haverkort, D. (2003). "Neural substrates and processing of a second language," in *The Lexicon–Syntax Interface in Second Language Acquisition*, eds R. van Hout, A. Hulk, F. Kuiken, and R. Towell (Amsterdam: John Benjamins), 151–174.

Sharwood Smith, M., and Truscott, J. (2011). "Consciousness and language: a processing perspective," in *New Horizons in the Neuroscience of Consciousness* eds E. Perry, E. D. Collerton, F. LeBeau, and H. Ashton (Amsterdam: John Benjamins), 129–138.

Sharwood Smith, M., and Truscott, J. (2014). *The Multilingual Mind: A Modular Processing Perspective*. Cambridge: Cambridge University Press.

Sharwood Smith, M., Truscott, J., and Hawkins, R. (2012). "Explaining change in transition grammars," in *A Handbook of Second Language Acquisition* eds J. Herschensohn and M. Young-Scholten (Cambridge: Cambridge University Press), 560–580.

Truscott, J., and Sharwood Smith, M. (2004). Acquisition by processing: a modular approach to language development. *Biling. Lang. Cogn.* 7, 1–20. doi: 10.1017/S1366728904001178

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: 31 May 2014; accepted: 20 June 2014; published online: 11 July 2014.

Citation: Sharwood Smith MA (2014) In search of conceptual frameworks for relating brain activity to language function. Front. Psychol. ***5****:716. doi: 10.3389/fpsyg.2014.00716*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

MINI REVIEW ARTICLE
published: 02 October 2014
doi: 10.3389/fpsyg.2014.01116

Structural brain changes related to bilingualism: does immersion make a difference?

***Maria Stein*[1,2,3] *, *Carmen Winkler*[1], *Anelis Kaiser*[2,4] *and Thomas Dierks*[1,3]**

[1] *Department of Psychiatric Neurophysiology, University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland*
[2] *Department of Clinical Psychology and Psychotherapy, Institute of Psychology, University of Bern, Bern, Switzerland*
[3] *Center for Cognition, Learning and Memory, University of Bern, Bern, Switzerland*
[4] *Department of Social Psychology, Institute of Psychology, University of Bern, Bern, Switzerland*

Edited by:
Christos Pliatsikas, University of Kent, UK

Reviewed by:
Ping Li, Penn State University, USA
Lorna Garcia Penton, Basque Center on Cognition, Brain and Language, Spain

****Correspondence:***
Maria Stein, Department of Psychiatric Neurophysiology, University Hospital of Psychiatry and Psychotherapy, University of Bern, Bolligenstrasse 111, 3000 Bern 60, Switzerland
e-mail: stein@puk.unibe.ch

Within the field of neuroscientific research on second language learning, considerable attention has been devoted to functional and recently also structural changes related to second language acquisition. The present literature review summarizes studies that investigated structural changes related to bilingualism. Furthermore, as recent evidence has suggested that native-like exposure to a second language (i.e., a naturalistic learning setting or immersion) considerably impacts second language learning, all findings are reflected with respect to the learning environment. Aggregating the existing evidence, we conclude that structural changes in left inferior frontal and inferior parietal regions have been observed in studies on cortical gray matter changes, while the anterior parts of the corpus callosum have been repeatedly found to reflect bilingualism in studies on white matter (WM) connectivity. Regarding the learning environment, no cortical alterations can be attributed specifically to naturalistic or classroom learning. With regard to WM changes, one might tentatively propose that changes in IFOF and SLF are possibly more prominently observed in studies investigating bilinguals with a naturalistic learning experience. However, future studies are needed to replicate and strengthen the existing evidence and to directly test the impact of naturalistic exposure on structural brain plasticity.

Keywords: bilingualism, naturalistic learning, immersion, second language, structural plasticity, VBM, DTI

INTRODUCTION

Experience-dependent changes in brain structure were first investigated in rodents placed in environmentally enriched versus very sparsely equipped standard cages. These early animal studies reported effects of environmental enrichment on brain weight (Rosenzweig et al., 1962) or cortical thickness (Rosenzweig et al., 1972) suggesting a structural adaptation process of the brain in response to experience. Since then, experience-dependent changes in human brain structure have been investigated in relation to various learning experiences, ranging from complex visuo-motor tasks like juggling to musical proficiency as well as to various aspects of language learning (see, e.g., May, 2011; Zatorre et al., 2012; Lovden et al., 2013).

A large proportion of the world's population is estimated to be bi- or multilingual (Bialystok, 2010; Grosjean and Li, 2013) and the importance of the ability to communicate in more than one language is even increasing in a globalized world. A growing body of literature has thus investigated this fascinating human ability and its neural underpinnings on a functional as well as on a structural level. Within this body of literature, studies targeting structural correlates of only one precisely defined language component (e.g., speech sound perception or grammatical skills) allow to postulate a direct relation between this very language domain and local brain structure (Golestani et al., 2007; Pliatsikas et al., 2014a). Other studies use a more global approach and relate bilingualism or a measure of global second language proficiency to brain structure. More specifically, in the first case, these studies simply compare brain structure of bilinguals to that of monolinguals. Potential differences are then related to bilingualism or general second language proficiency. In the second case, these studies typically assess second language proficiency either by a variety of language tests or by overall scores such as school grades and relate these measures to brain structure. In the context of the present research topic on naturalistic exposure, this Mini-Review will summarize these studies on global L2-learning with a special focus on the question whether certain change patterns can be related to the environment in which the second language has been learned (naturalistic learning through immersion vs. classroom setting). Naturalistic language learning through immersion is characterized by high levels of L2-exposure and implicit learning; it is thus similar to L1 acquisition. In contrast, traditional L2 classroom instruction is mainly based on formalized training exercises and explicit instruction (e.g., Dahl and Vulchanova, 2014). Due to these differences, it is thus well conceivable that the structural change patterns evoked by these two learning types differ.

SELECTION PROCESS FOR INCLUSION OF STUDIES IN THIS MINI-REVIEW

A systematic literature search was conducted in the databases PubMed/MEDLINE and Google scholar in April 2014. After having read through the resulting literature, appropriate studies were selected according to the following criteria: (1) published in a peer reviewed journal, (2) brain regions specifically associated with

overall L2-acquisition, (3) original research articles, and (4) concentration exclusively on global language proficiency (vs. a single language domain). As bilinguals per definition have higher L2 proficiency than monolinguals, group comparisons between mono- and bilinguals were also considered as meeting this last criterion. Studies on aging effects were not considered in this Mini review. While 27 studies met criteria 1–3, a subgroup of eleven concentrated on global language proficiency. One of these, a post-mortem single case study (Amunts et al., 2004), differed significantly in methodology and was therefore excluded. Ultimately 10 research papers met all four criteria and were included in this Mini-Review (see Li et al., 2014 for a more extensive review). All studies investigated structural brain changes related to overall second language proficiency. In the following, the studies will be grouped according to their focus on either gray matter (GM; **Table 1**) or white matter (WM) changes (**Table 2**) and will be discussed with respect to main findings and inferences that might be drawn regarding immersion into a second language.

GRAY MATTER CHANGES RELATED TO SECOND LANGUAGE LEARNING

A seminal study on structural changes related to L2-acquisition compared early (age of acquisition (AoA) < 5 years) and late bilinguals (AoA: 10–15 years) to English monolinguals (Mechelli et al., 2004). Gray matter density (GMD) was higher in the left inferior parietal cortex (l-IPC) in bilinguals compared to monolinguals. This increase was even more pronounced in early relative to late bilinguals. In a second sample of Italian-English bilinguals (AoA: 2–34 years) the increase in L-IPC correlated positively with the degree of L2-proficiency and negatively with AoA. Mechelli et al. (2004) thus presented first strong evidence for structural changes related to bilingualism. However, with respect to the cross-sectional design of the study, it remained unclear whether the observed changes were directly induced by the experience of learning another language. The first study approaching this question in a longitudinal design (Osterhout et al., 2008) measured MRI in four University students enrolled in a 9 week intensive (3.5 h/day) Spanish course at two points in time (at the beginning and the end of the course). Because of the small sample size, the authors conducted a region-of-interest analysis in l-IPC and reported increasing GMD over the course of L2-acquisition, suggesting that structural changes in l-IPC are experience-dependent.

In a study with very high immersion to a L2-environment, native English-speaking exchange students learning German in Switzerland participated in language proficiency tests and MRI-measurements once at the beginning of their stay and a second time about 5 months later (Stein et al., 2012). The individual amount of learning (i.e., the L2-test score difference between first and second measurement) correlated with the increase in GMD in the left inferior frontal gyrus (l-IFG) as well as in the left anterior temporal lobe (l-ATL). While additional analyses exploring the effects of maturation and general environmental enrichment could not rule out the possibility that the l-ATL-cluster is due to these effects, the l-IFG-cluster seemed to be specifically linked to increasing L2-proficiency. The l-IFG changes thus reflected the individual amount of L2- learning (regardless of absolute proficiency). Martensson et al. (2012) investigated L2-acquisition through intense classroom-instruction and examined conscripts in the interpreter academy of the Swedish military, where a new language is learned to fluency within 10 months. These interpreters and monolingual controls participated in MRI-measurement immediately before the interpreter academy started and 3 months later. The grade on the mid-year exam (taken a few weeks after the second MRI-measurement) served as an indicator for language proficiency. Compared to controls, interpreters displayed larger pre-to-post-increases in cortical thickness in left middle frontal gyrus (l-MFG), l-IFG, left superior temporal gyrus (l-STG) as well as larger increases in hippocampi volumes. Furthermore, changes in right hippocampus and l-STG cortical thickness correlated with L2-proficiency level. Klein et al. (2014) compared cortical thickness of simultaneous bilinguals (AoA 0–3 years), early sequential bilinguals (AoA 4–7 years), and late sequential bilinguals (AoA 8–13 years) to monolingual controls. Interestingly, they observed differences in cortical thickness in l-IFG (higher thickness in bilinguals) and r-IFG (lower thickness in bilinguals) when comparing early and late bilinguals to monolingual controls, while no brain region differed significantly between simultaneous bilinguals and monolingual controls. Comparing Spanisch-Catalan bilinguals to Spanish monolinguals, Ressel et al. (2012) found no significant differences in a whole-brain VBM-analysis, but observed larger bilateral Heschl's gyrus in the bilinguals.

WHITE MATTER CHANGES RELATED TO SECOND LANGUAGE LEARNING

Most studies on WM changes used diffusion tensor imaging (DTI) to measure amount and directionality of water diffusion. When this diffusion is restricted in one direction more than in another (e.g., by an axonal cell membrane), water diffusion becomes anisotopic. Fractional anisotropy (FA) indicates to which degree water diffusivity is unimpeded (low FA) or restricted (high FA). FA values are typically higher along axonal bundles and thus allow investigating connectivity in the human brain (Conturo et al., 1999; Le Bihan et al., 2001).

To observe how bilingualism impacts WM pathways, Mohades et al. (2012) recruited simultaneous bilingual children (L2-exposure since birth), sequential bilingual children (AoA > 3 years) and monolingual children. Mean FA was assessed in four selected tracts with relevance to language processing. Higher FA-values were found in the left inferior frontal-occipital fasciculus (l-IFOF) in simultaneous bilinguals compared to sequential bilinguals and monolinguals. Furthermore, in the bundle arising from anterior corpus callosum (CC) and projecting into the orbital lobe (AC-OL) simultaneous bilinguals displayed lower FA-values than monolinguals. While the l-IFOF-finding suggests faster transmission of semantic information in simultaneous bilinguals (Duffau et al., 2005; Mandonnet et al., 2007), the interpretation of the AC-OL finding remains unclear (Mohades et al., 2012). Coggins et al. (2004) analyzed the variability of CC and compared the relative size of CC-subregions in bilinguals compared to monolinguals. The authors observed larger relative anterior-midbody CC in bilinguals and interpret this in the light of greater processing demands of multiple languages which require

Table 1 | Overview of studies investigating gray matter changes related to global second language proficiency.

Author	Sample	Learning environment	Method	Analyses	Main results
Mechelli et al. (2004)	(A) 25 early and 33 late bilinguals; 25 monolingual controls	Early bilinguals: naturalistic setting Late bilinguals: mixed (class-room setting for some, naturalistic learning for others)	VBM (GMD)	Cross-sectional group comparisons (bilingual vs. monolinguals, whole-brain approach)	l-IPC: bilinguals > monolinguals
	(B) 22 bilinguals	Unknown	VBM (GMD)	Correlation of VBM-changes with proficiency and AOA in bilinguals (whole-brain approach)	l-IPC: positive correlation with L2-proficiency l-IPC: negative correlation with L2-AOA
Osterhout et al. (2008)	Four students learning Spanish in a university course	Non-naturalistic, classroom setting	VBM (GMD)	Pre–post-comparison in l-IPC (ROI-approach)	l-IPC: increases from pre to post
Stein et al. (2012)	10 exchange students learning German	Naturalistic setting	VBM (GMD)	Correlation of VBM-changes with proficiency-changes (whole-brain approach)	GMD-increase in l-IFG correlates with individual increase in proficiency
Martensson et al. (2012)	14 conscripts in the interpreter academy; 17 monolingual controls	Non-naturalistic setting	Cortical thickness, subcortical gray matter volume	Group comparison (interpreters vs. controls, whole brain approach) of cortical thickness and subcortical gray matter changes (pre–post) Correlation of brain changes with proficiency changes (ROI-approach based on group comparison)	Interpreter (vs. controls) showed higher increase in cortical thickness (pre–post) in l-MFG, l-IFG, l-STG and in bilateral hippocampal volume Proficiency correlated with increase in r hippocampus and l-STG
Ressel et al. (2012)	22 Catalan-Spanish bilinguals; 22 Spanish monolinguals	Naturalistic setting	VBM; volumetric measurement of HG	Group comparison (bilinguals vs. monolinguals, whole-brain brain approach) of VBM values. Group comparison of manually segmented HG volumes	No VBM differences at corrected threshold. Bilinguals had higher HG volumes than monolinguals
Klein et al. (2014)	22 simultaneous bilinguals; 22 early sequential bilinguals; 22 late sequential bilinguals; 22 monolinguals	Naturalistic setting	Cortical thickness	Cross-sectional group comparisons (bilingual vs. monolinguals)	Cortical thickness in l-IFG late bilingual > monolingual and early bilingual > monolingual Cortical thickness in r-IFG in monolingual > late bilingual, monolingual > early bilingual Simultaneous bilingual > late bilingual and early bilingual > late bilingual

VBM, voxel-based morphometry; GMD, gray matter density; l, left; r, right; IPC, inferior parietal cortex; IFG, inferior frontal gyrus; MFG, middle frontal gyrus; STG, superior temporal gyrus; HG, Heschl's gyrus; ROI, region of interest.

Table 2 | Overview of studies investigating white matter changes related to global second language proficiency.

Author	Sample	Learning environment	Analyses	Method	Main results
Coggins et al. (2004)	12 Bilinguals (seven early bilinguals; five late bilinguals); seven monolinguals	Classroom setting	Compare CC morphology (regional to total area ratio) between groups	Analyses of size (regional to total area ratio) of five CC subregions as defined on the midsagittal plane of an MRI image	AMB ratio to total CC: larger in bilinguals than in monolinguals
Mohades et al. (2012)	Children (8–11 years): 15 simultaneous bilinguals; 15 sequential bilinguals; 15 monolinguals	Simultaneous bilinguals: naturalistic Sequential bilinguals: classroom-setting	Group comparison of four preselected white matter tracts (AF/SLF, IFOF, AC-OL, AMB)	DTI	Left IFOF: FA in simultaneous bilinguals > sequential bilinguals > monolinguals AC-OL: FA in simultaneous bilinguals < sequential bilinguals < monolinguals
Schlegel et al. (2012)	11 English speaking students learning Chinese; 16 monolingual controls	Classroom setting	Longitudinal study comparing DTI changes in second language learners to those of mono-linguals	DTI (FA and RD)	Progressive FA increase in second language learners > controls FA increase in tracts connecting bilateral IFG, FMG, FPG, CN, left STG, PP, right PT FA increase related to second language proficiency
García-Pentón et al. (2014)	13 Spanish monolinguals; 13 Spanish-Basque bilinguals	Naturalistic	Group comparison of connectivity differences (whole brain approach)	DTI-based connectivity analysis, network based statistics, graph analysis	Two networks show higher connectivity and more graph-efficient information flow: (a) left-sided network comprising SMG, STG, IFG, MSFG, INS. (b) network comprising right SFG, left SPG, ANG, STP, SOG

CC, corpus callosum; IFOF, inferior occipitofrontal fasciculus; AF, arcuate fasciculus; SLF, superior longitudinal fasciculus; AC-OL, connection anterior corpus callosum and orbital lobes; AMB, anterior-midbody of corpus callosum; UNF, uncinated fasciculus; FMG, frontomarginal gyrus; FPG, frontopolar gyrus; CN, caudate nucleus; PP, planum polare; PT, planum temporale; SMG, supramarginal gyrus; STG, superior temporal gyrus; IFG, inferior frontal gyrus; (M)SFG, (medial) superior frontal gyrus; INS, insula; SPG, superior parietal gyrus; ANG, angular gyrus; STP, superior temporal gyrus; SOG, superior occipital gyrus; DTI, diffusion tensor imaging; FA, fractional anisotropy; RD, radial diffusivity.

stronger interhemispheric communication of cortical regions bridged by anterior-midbody CC.

A recent study by García-Pentón et al. (2014) investigated anatomical connectivity in early Spanish-Basque bilinguals and native Spanish monolingual controls using a DTI-based tractography technique and network-based statistics. Bilinguals displayed increased connectivity in two networks: One left hemispheric network connecting frontal, parietal and temporal regions, and one network involving left occipital, temporal and parietal regions as well as right superior frontal gyrus. Within these networks, a graph-analytic approach indicated that in addition to higher connectivity, there is also more efficient information flow.

While the studies above all represent group comparisons investigating long term changes in WM, Schlegel et al. (2012) opted for a more dynamic approach to observe WM-changes in adults learning a new language. In this longitudinal study, monthly DTI scans were collected from English speaking students enrolled in a 9 month intensive Chinese course and from controls. Chinese-learners displayed a significant increase in connectivity (measured as increased FA and decreased radial diffusivity)

in a network connecting left hemisphere language regions and their right hemisphere analogs (e.g., IFG, caudate nucleus, STG). The most prominent changes occurred in the frontal tracts that cross the anterior-CC, speaking for an increased interhemispheric connectivity in Chinese-learners. The authors also show that FA increases progressively over time and that this increase is related to the level of second language proficiency.

DISCUSSION

Even if the evidence is still sparse and considerable differences between studies exist, an aggregation of the findings on **gray matter changes** suggests that structural changes in l-IPC and l-IFG seem to be most consistently related to measures of global second language learning or bilingualism. Both regions have also been repeatedly linked to second language proficiency in studies on functional brain activation (e.g., Chee et al., 2001; Perani et al., 2003; Sakai et al., 2004; Stein et al., 2006, 2009; Raboyeau et al., 2010; Li et al., 2014).

Concerning the learning environment, l-IPC was observed to vary with L2-learning irrespective of the learning setting: The early bilinguals as well as part of the late bilinguals in the Mechelli-study learned L2 through naturalistic exposure while at least another group of the late bilinguals in the Mechelli-study acquired L2 through classroom instruction (personal communication Cathy Price, June 3rd 2014, Andrea Mechelli, June 4th 2014), the latter being also true for participants in the study by Osterhout et al. (2008). Despite this variation, l-IPC-changes have been observed in all three groups, suggesting that these changes seem to accompany L2-acquisition irrespective of learning setting. Even if the l-IPC-changes are more pronounced in early compared to late bilinguals, this variation is most likely attributed to differences in AoA (Mechelli et al., 2004).

Concerning the influence of naturalistic immersion on the l-IFG finding, a group with extensive immersion into an L2-environment (exchange students in Stein et al., 2012) as well as the group with the clearest non-naturalistic, classroom setting (interpreters in Martensson et al., 2012) displayed structural changes in l-IFG. In turn, the early Spanish-Catalan bilinguals in Ressel et al.'s (2012) study did not display l-IFG changes. One might argue that the results of Ressel et al. (2012; where the only difference was observed in bilateral HG) might also be due to the fact that the bilinguals' two languages mainly differed in phonology, while having a considerable lexical overlap. On the other hand, together with the study by Klein et al. (2014), where the simultaneous bilinguals were the only bilingual group without increased l-IFG thickness, this might also to suggest that it is not immersion as such but rather "age of immersion" that might influence whether structural l-IFG changes occur or not: As the l-IFG is involved in cognitive language control (Abutalebi and Green, 2008; Luk et al., 2011b), controlled retrieval (Rodriguez-Fornells et al., 2009) and morphosyntactic processing (Pliatsikas et al., 2014b), it might be particularly recruited by explicit learning. This learning type, even if directly targeted only by traditional classroom instruction, might also be deployed by late bilinguals during highly immersed learning. Such an interpretation is in line with the observation that younger learners outperform older ones in implicit learning while older learners are more apt to rely on (and better in) explicit learning (DeKeyser and Larson-Hall, 2005; Muñoz, 2006).

Regarding other brain regions reported to vary with global second language proficiency, l-STG and l-MFG were until now only observed to change in a classroom setting (Martensson et al., 2012). The fact that the second study with a classroom setting did not replicate these findings must not be mistaken as conflicting evidence, as Osterhout et al. (2008) only performed a ROI-analysis of l-IPC. When extending the focus to studies on single L2 domains (like, e.g., Li et al., 2014), the only study that observed structural changes in these regions when analyzing L2-learning in a naturalistic setting was conducted by Crinion et al. (2009) and related l-STG changes to the acquisition of a tonal as opposed to non-tonal languages. However, in both regions differential functional activation in L2-processing was also observed in samples with high L2-immersion (e.g., Parker Jones et al., 2012; Archila-Suerte et al., 2013). Furthermore, a study on L2-related WM changes in a highly immersed sample (García-Pentón et al., 2014) reported connectivity changes in a network including STG. In the case of STG, its assumed functional role in phonological processing (Callan et al., 2004; Zheng et al., 2010), additionally undermines the assumption that naturalistic L2 learning should be less effective in inducing STG changes. It thus seems likely that the failure to observe structural changes in these regions in response to naturalistic L2 learning is merely due a lack of research on that issue. Thus, future studies directly comparing naturalistic learning and classroom instruction while controlling for differences in AoA and proficiency level are necessary to determine the influence of L2-immersion on gray matter changes.

Taken together, the studies on **WM changes** repeatedly reported L2-related changes in the CC, the main anatomical link between left and right hemispheres. Generally, this finding is in line with the observation that the language network in bilinguals seems to be less left lateralized and more bilateral compared to monolinguals (Hull and Vaid, 2006), while the compatibility with the assumption that AoA is the most important factor in determining the degree of lateralization (Hull and Vaid, 2007) seems less clear. Note that most of the data on CC-changes stems from group comparisons between bi- and monolinguals, making it hard to draw inferences about the dynamic of these changes as well as about the role of the precise proficiency level. Regarding the precise portion of the CC that adapts in response to L2-exposure, the anterior and anterior-midbody CC seem to be candidate regions: Coggins reported relative larger anterior-midbody CC, Mohades et al. (2012) observed lower FA-values in the anterior part of CC in bilingual children, and a large part of the regions found to be increasingly interconnected in the study by Schlegel et al. (2012) are anatomically connected via anterior to mid-CC. However, there are still inconsistencies and open questions regarding the factors (e.g., age) influencing structural changes in CC. Furthermore, both, changes in relative volume (Coggins et al., 2004) as well as FA-changes (Mohades et al., 2012) may be due to different axonal characteristics [e.g., myelination, axonal density, axonal caliber, and fiber coherence (Cheng et al., 2010)], thus the precise nature of the underlying adaptation remain to be further explored. Considering the effects of naturalistic exposure, the CC

seems to undergo changes in response to naturalistic L2 learning (simultaneous bilinguals in Mohades et al., 2012) as well as during classroom instruction (Coggins et al., 2004; Schlegel et al., 2012).

Another fiber tract adapting when people acquire a second language seems to be the IFOF (Mohades et al., 2012). These results are in line with studies relating the IFOF to semantic processing (e.g., Duffau, 2008; Martino et al., 2010) as well as with studies on WM integrity in elderly bilinguals (e.g., Luk et al., 2011a; but see Gold et al., 2013; for changes in the opposite direction). Very interestingly, when looking closely at the results, IFOF-changes might be most pronounced when L2 is learned through naturalistic exposure: The group with the most pronounced IFOF-effects in the study by Mohades et al. (2012) acquired their L2 through naturalistic exposure and so did the elderly sample in the Luk et al. (2011a) study. Not in line with this interpretation, however, is the absence of IFOF changes in the study by García-Pentón et al. (2014), which examined bilinguals with high levels of L2-immersion. In turn, García-Pentón et al. (2014) reported increased connectivity in a left-sided network comprising frontal regions as well as supramarginal gyrus, thus a network that is partly connected via the superior longitudinal fasciculus (SLF). Consistently, the elderly bilinguals with naturalistic exposure in Luk et al. (2011a) equally displayed SLF-alterations. This might indicate that immersion in L2 (in contrast to pure classroom instruction) has a stronger influence on SLF-changes. A study directly comparing two bilingual groups with different learning experiences however failed to find SLF differences (Mohades et al., 2012). Thus, future studies are needed to enlighten the effects of immersion on WM changes.

REFERENCES

Abutalebi, J., and Green, D. W. (2008). Control mechanisms in bilingual language production: neural evidence from language switching studies. *Lang. Cogn. Process.* 23, 557–582. doi: 10.1080/01690960801920602

Amunts, K., Schleicher, A., and Zilles, K. (2004). Outstanding language competence and cytoarchitecture in Broca's speech region. *Brain Lang.* 89, 346–353. doi: 10.1016/S0093-934X(03)00360-2

Archila-Suerte, P., Zevin, J., Ramos, A. I., and Hernandez, A. E. (2013). The neural basis of non-native speech perception in bilingual children. *Neuroimage* 67, 51–63. doi: 10.1016/j.neuroimage.2012.10.023

Bialystok, E. (2010). Bilingualism. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 559–572. doi: 10.1002/Wcs.43

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. doi: 10.1016/j.neuroimage.2004.03.006

Chee, M. W., Hon, N., Lee, H. L., and Soon, C. S. (2001). Relative language proficiency modulates BOLD signal change when bilinguals perform semantic judgments. Blood oxygen level dependent. *Neuroimage* 13, 1155–1163. doi: 10.1006/nimg.2001.0781

Cheng, Y., Chou, K. H., Chen, I. Y., Fan, Y. T., Decety, J., and Lin, C. P. (2010). Atypical development of white matter microstructure in adolescents with autism spectrum disorders. *Neuroimage* 50, 873–882. doi: 10.1016/j.neuroimage.2010.01.011

Coggins, P. E. III, Kennedy, T. J., and Armstrong, T. A. (2004). Bilingual corpus callosum variability. *Brain Lang.* 89, 69–75. doi: 10.1016/S0093-934X(03)00299-2

Conturo, T. E., Lori, N. F., Cull, T. S., Akbudak, E., Snyder, A. Z., Shimony, J. S., et al. (1999). Tracking neuronal fiber pathways in the living human brain. *Proc. Natl. Acad. Sci. U.S.A.* 96, 10422–10427. doi: 10.1073/pnas.96.18.10422

Crinion, J. T., Green, D. W., Chung, R., Ali, N., Grogan, A., Price, G. R., et al. (2009). Neuroanatomical markers of speaking Chinese. *Hum. Brain Mapp.* 30, 4108–4115. doi: 10.1002/hbm.20832

Dahl, A., and Vulchanova, M. D. (2014). Naturalistic acquisition in an early language classroom. *Front. Psychol.* 5:329. doi: 10.3389/fpsyg.2014.00329

DeKeyser, R., and Larson-Hall, J. (2005). "What does the critical period really mean?," in *Handbook of Bilingualism: Psycholinguistic Approaches*, eds J. F. Kroll and A. M. B. De Groot (New York, NY: Oxford University Press).

Duffau, H. (2008). The anatomo-functional connectivity of language revisited. New insights provided by electrostimulation and tractography. *Neuropsychologia* 46, 927–934. doi: 10.1016/j.neuropsychologia.2007.10.025

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. doi: 10.1093/brain/awh423

García-Pentón, L., Perez Fernandez, A., Iturria-Medina, Y., Gillon-Dowens, M., and Carreiras, M. (2014). Anatomical connectivity changes in the bilingual brain. *Neuroimage* 84, 495–504. doi: 10.1016/j.neuroimage.2013.08.064

Gold, B. T., Johnson, N. F., and Powell, D. K. (2013). Lifelong bilingualism contributes to cognitive reserve against white matter integrity declines in aging. *Neuropsychologia* 51, 2841–2846. doi: 10.1016/j.neuropsychologia.2013.09.037

Golestani, N., Molko, N., Dehaene, S., Lebihan, D., and Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cereb. Cortex* 17, 575–582. doi: 10.1093/cercor/bhk001

Grosjean, F., and Li, P. (2013). *The Psycholinguistics of Bilingualism.* New York, NY: Wiley-Blackwell.

Hull, R., and Vaid, J. (2006). Laterality and language experience. *Laterality* 11, 436–464. doi: 10.1080/13576500600691162

Hull, R., and Vaid, J. (2007). Bilingual language lateralization: a meta-analytic tale of two hemispheres. *Neuropsychologia* 45, 1987–2008. doi: 10.1016/j.neuropsychologia.2007.03.002

Klein, D., Mok, K., Chen, J. K., and Watkins, K. E. (2014). Age of language learning shapes brain structure: a cortical thickness study of bilingual and monolingual individuals. *Brain Lang.* 131, 20–24. doi: 10.1016/j.bandl.2013.05.014

Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., et al. (2001). Diffusion tensor imaging: concepts and applications. *J. Magn. Reson. Imaging* 13, 534–546. doi: 10.1002/Jmri.1076

Li, P., Legault, J., and Litcofsky, K. A. (2014). Neuroplasticity as a function of second language learning: anatomical changes in the human brain. *Cortex* 58, 301–324. doi: 10.1016/j.cortex.2014.05.001

Lovden, M., Wenger, E., Martensson, J., Lindenberger, U., and Backman, L. (2013). Structural brain plasticity in adult learning and development. *Neurosci. Biobehav. Rev.* 37, 2296–2310. doi: 10.1016/j.neubiorev.2013.02.014

Luk, G., Bialystok, E., Craik, F. I., and Grady, C. L. (2011a). Lifelong bilingualism maintains white matter integrity in older adults. *J. Neurosci.* 31, 16808–16813. doi: 10.1523/JNEUROSCI.4563-11.2011

Luk, G., Green, D. W., Abutalebi, J., and Grady, C. (2011b). Cognitive control for language switching in bilinguals: a quantitative meta-analysis of functional neuroimaging studies. *Lang. Cogn. Process.* 27, 1479–1488. doi: 10.1080/01690965.2011.613209

Mandonnet, E., Nouet, A., Gatignol, P., Capelle, L., and Duffau, H. (2007). Does the left inferior longitudinal fasciculus play a role in language? A brain stimulation study. *Brain* 130, 623–629. doi: 10.1093/brain/awl361

Martensson, J., Eriksson, J., Bodammer, N. C., Lindgren, M., Johansson, M., Nyberg, L., et al. (2012). Growth of language-related brain areas after foreign language learning. *Neuroimage* 63, 240–244. doi: 10.1016/j.neuroimage.2012.06.043

Martino, J., Brogna, C., Robles, S. G., Vergani, F., and Duffau, H. (2010). Anatomic dissection of the inferior fronto-occipital fasciculus revisited in the lights of brain stimulation data. *Cortex* 46, 691–699. doi: 10.1016/j.cortex.2009.07.015

May, A. (2011). Experience-dependent structural plasticity in the adult human brain. *Trends Cogn. Sci.* 15, 475–482. doi: 10.1016/j.tics.2011.08.002

Mechelli, A., Crinion, J. T., Noppeney, U., O'Doherty, J., Ashburner, J., Frackowiak, R. S., et al. (2004). Neurolinguistics: structural plasticity in the bilingual brain. *Nature* 431, 757. doi: 10.1038/431757a

Mohades, S. G., Struys, E., Van Schuerbeek, P., Mondt, K., Van De Craen, P., and Luypaert, R. (2012). DTI reveals structural differences in white matter tracts between bilingual and monolingual children. *Brain Res.* 1435, 72–80. doi: 10.1016/j.brainres.2011.12.005

Muñoz, C. (2006). "The effects of age on foreign language learning: the BAF project," in *Age and the Rate of Foreign Language Learning*, ed. C. Munoz (Great Britain: Cromwell Press), 1–40.

Osterhout, L., Poliakov, A., Inoue, K., Mclaughlin, J., Valentine, G., Pitkanen, I., et al. (2008). Second-language learning and changes in the brain. *J. Neurolinguistics* 21, 509–521. doi: 10.1016/j.jneuroling.2008.01.001

Parker Jones, O., Green, D. W., Grogan, A., Pliatsikas, C., Filippopolitis, K., Ali, N., et al. (2012). Where, when and why brain activation differs for bilinguals and monolinguals during picture naming and reading aloud. *Cereb. Cortex* 22, 892–902. doi: 10.1093/cercor/bhr161

Perani, D., Abutalebi, J., Paulesu, E., Brambati, S., Scifo, P., Cappa, S. F., et al. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: an fMRI study during verbal fluency. *Hum. Brain Mapp.* 19, 170–182. doi: 10.1002/hbm.10110

Pliatsikas, C., Johnstone, T., and Marinis, T. (2014a). Grey matter volume in the cerebellum is related to the processing of grammatical rules in a second language: a structural voxel-based morphometry study. *Cerebellum* 13, 55–63. doi: 10.1007/s12311-013-0515-6

Pliatsikas, C., Johnstone, T., and Marinis, T. (2014b). FMRI evidence for the involvement of the procedural memory system in morphological processing of a second language. *PLoS ONE* 9:e97298. doi: 10.1371/journal.pone.0097298

Raboyeau, G., Marcotte, K., Adrover-Roig, D., and Ansaldo, A. I. (2010). Brain activation and lexical learning: the impact of learning phase and word type. *Neuroimage* 49, 2850–2861. doi: 10.1016/j.neuroimage.2009.10.007

Ressel, V., Pallier, C., Ventura-Campos, N., Díaz, B., Roessler, A., Ávila, C., et al. (2012). An effect of bilingualism on the auditory cortex. *J. Neurosci.* 32, 16597–16601. doi: 10.1523/JNEUROSCI.1996-12.2012

Rodriguez-Fornells, A., Cunillera, T., Mestres-Misse, A., and De Diego-Balaguer, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 3711–3735. doi: 10.1098/rstb.2009.0130

Rosenzweig, M. R., Bennett, E. L., and Diamond, M. C. (1972). Brain changes in response to experience. *Sci. Am.* 226, 22–29. doi: 10.1038/scientificamerican0272-22

Rosenzweig, M. R., Krech, D., Bennett, E. L., and Zolman, J. F. (1962). Variation in environmental complexity and brain measures. *J. Comp. Physiol. Psychol.* 55, 1092–1095. doi: 10.1037/h0042758

Sakai, K. L., Miura, K., Narafu, N., and Muraishi, Y. (2004). Correlated functional changes of the prefrontal cortex in twins induced by classroom education of second language. *Cereb. Cortex* 14, 1233–1239. doi: 10.1093/cercor/bhh084

Schlegel, A. A., Rudelson, J. J., and Tse, P. U. (2012). White matter structure changes as adults learn a second language. *J. Cogn. Neurosci.* 24, 1664–1670. doi: 10.1162/jocn_a_00240

Stein, M., Dierks, T., Brandeis, D., Wirth, M., Strik, W., and Koenig, T. (2006). Plasticity in the adult language system: a longitudinal electrophysiological study on second language learning. *Neuroimage* 33, 774–783. doi: 10.1016/j.neuroimage.2006.07.008

Stein, M., Federspiel, A., Koenig, T., Wirth, M., Lehmann, C., Wiest, R., et al. (2009). Reduced frontal activation with increasing 2nd language proficiency. *Neuropsychologia* 47, 2712–2720. doi: 10.1016/j.neuropsychologia.2009.05.023

Stein, M., Federspiel, A., Koenig, T., Wirth, M., Strik, W., Wiest, R., et al. (2012). Structural plasticity in the language system related to increased second language proficiency. *Cortex* 48, 458–465. doi: 10.1016/j.cortex.2010.10.007

Zatorre, R. J., Fields, R. D., and Johansen-Berg, H. (2012). Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nat. Neurosci.* 15, 528–536. doi: 10.1038/nn.3045

Zheng, Z. Z., Munhall, K. G., and Johnsrude, I. S. (2010). Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production. *J. Cogn. Neurosci.* 22, 1770–1781. doi: 10.1162/jocn.2009.21324

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: 20 June 2014; accepted: 15 September 2014; published online: 02 October 2014.

Citation: Stein M, Winkler C, Kaiser A and Dierks T (2014) Structural brain changes related to bilingualism: does immersion make a difference? Front. Psychol. ***5****:1116. doi: 10.3389/fpsyg.2014.01116*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 June 2015
doi: 10.3389/fpsyg.2015.00638

Age of second language acquisition in multilinguals has an impact on gray matter volume in language-associated brain areas

Anelis Kaiser[1]†, Leila S. Eppenberger[2†], Renata Smieskova[3], Stefan Borgwardt[3], Esther Kuenzli[4], Ernst-Wilhelm Radue[2], Cordula Nitsch[5‡] and Kerstin Bendfeldt[2‡]*

[1] Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern, Bern, Switzerland, [2] Medical Image Analysis Centre, University Hospital Basel, Basel, Switzerland, [3] Department of Psychiatry, University Hospital Basel, University of Basel, Basel, Switzerland, [4] Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland, [5] Department of Biomedicine, Institute of Anatomy, University of Basel, Basel, Switzerland

OPEN ACCESS

Edited by:
Vicky Chondrogianni, University of Edinburgh, UK

Reviewed by:
Judith F. Kroll, The Pennsylvania State University, USA
Christos Pliatsikas, University of Kent, UK

***Correspondence:**
Anelis Kaiser, Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland
anelis.kaiser@psy.unibe.ch

† These authors have shared first authorship.
‡ These authors have shared last authorship.

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: 16 June 2014
Accepted: 01 May 2015
Published: 08 June 2015

Citation:
Kaiser A, Eppenberger LS, Smieskova R, Borgwardt S, Kuenzli E, Radue E-W, Nitsch C and Bendfeldt K (2015) Age of second language acquisition in multilinguals has an impact on gray matter volume in language-associated brain areas. Front. Psychol. 6:638. doi: 10.3389/fpsyg.2015.00638

Numerous structural studies have established that experience shapes and reshapes the brain throughout a lifetime. The impact of early development, however, is still a matter of debate. Further clues may come from studying multilinguals who acquired their second language at different ages. We investigated adult multilinguals who spoke three languages fluently, where the third language was learned in classroom settings, not before the age of 9 years. Multilinguals exposed to two languages simultaneously from birth (SiM) were contrasted with multinguals who acquired their first two languages successively (SuM). Whole brain voxel based morphometry revealed that, relative to SuM, SiM have significantly lower gray matter volume in several language-associated cortical areas in both hemispheres: bilaterally in medial and inferior frontal gyrus, in the right medial temporal gyrus and inferior posterior parietal gyrus, as well as in the left inferior temporal gyrus. Thus, as shown by others, successive language learning increases the volume of language-associated cortical areas. In brains exposed early on and simultaneously to more than one language, however, learning of additional languages seems to have less impact. We conclude that – at least with respect to language acquisition – early developmental influences are maintained and have an effect on experience-dependent plasticity well into adulthood.

Keywords: multilingualism, bilingualism, age of L2 acquisition, magnetic resonance imaging, gray matter volume

Introduction

In recent years, numerous studies on neuronal plasticity have established that training results in structural changes in critically involved cortical brain areas. On a macroscopic level, it has been shown that gray matter (GM) density and GM volume are altered after different kinds of training (for review see Taubert et al., 2010 and Zatorre et al., 2012). Within the domain of neurolinguistics too, ongoing research has demonstrated that acquiring a second language (L2) has a substantial influence on the anatomy of the brain (Li et al., 2014; Stein et al., 2014). This was the case for a variety of language characteristics, such as non-native speech sounds (Golestani et al., 2002, 2007, 2011; Crinion et al., 2006; Wong et al., 2008), acquisition of vocabulary (Grogan et al., 2009, 2012; Hosoda et al., 2013), reading skills

(Cummine and Boliek, 2013; Zhang et al., 2013), syntax abilities (Nauchi and Sakai, 2009; Pliatsikas et al., 2014) and executive language control (Elmer et al., 2011; Filippi et al., 2011; Abutalebi et al., 2012, 2013b; Zou et al., 2012). In addition to these studies on specific characteristics of L2 acquisition, research has also been devoted to overall second language proficiency (Amunts et al., 2004; Coggins et al., 2004; Mechelli et al., 2004; Osterhout et al., 2008; Mårtensson et al., 2012; Mohades et al., 2012; Ressel et al., 2012; Schlegel et al., 2012; Stein et al., 2012; García-Pentón et al., 2014; Klein et al., 2014), i.e., to a broad level of L2 proficiency as assessed by overall linguistic testing or by simply comparing groups of bilinguals to groups of monolinguals (Stein et al., 2014; Winkler, unpublished master thesis). Anatomical studies in bilinguals are based on the investigation of GM density changes (e.g., Mechelli et al., 2004; Osterhout et al., 2008; Stein et al., 2012), cortical thickness (e.g., Mårtensson et al., 2012; Klein et al., 2014), and GM volume (e.g., Golestani et al., 2007; Wong et al., 2008). In language-associated areas and, in particular, in areas implicated in control, these studies have generally found alterations due to L2 acquisition, that increase with growing L2 proficiency (for a review, see Li et al., 2014; Stein et al., 2014). Most recently, there have been morphometric studies suggesting that the consequences of bilingualism are related to the form of bilingualism, including not only the age of acquisition and proficiency, but also the context in which the two or more languages are used, i.e., whether speakers are immersed in the language environment, whether they learned the language in a classroom setting and so forth (see the current Special issue).

Neuroscientific studies with multilingual participants, i.e., with subjects speaking at least three languages fluently, are not carried out very often. Only recently, Abutalebi et al. (2013a) combined functional and structural MRI to examine the role of the basal ganglia in multilingual participants. Because of the permanent exposure to a major articulatory load when speaking several languages during a lifetime, the authors hypothesized that there is an enlarged density of gray matter in this area in multilinguals. Indeed, as compared to monolinguals, participants speaking three languages demonstrated increased GM density in the left putamen, which supports the notion of structural plasticity as result of handling a complex articulatory repertoire. In order to compare high cognitive, linguistic, and articulatory demands between multilinguals of two different sorts, namely professional multilingual interpreters versus control multilinguals, Elmer et al. (2014) conducted GM volume analysis on regions previously shown to support language control and executive functions in multilinguals. Interestingly, GM volume was found to be reduced in highly trained multilingual interpreters in a number of regions associated with language control, which suggests that intense training can result in more efficient neural networks, probably due to the pruning of superfluous connections.

Most of the research on the (co)-organization of several languages in the brains of multilinguals has been conducted by using functional brain imaging. Vingerhoets et al. (2003) demonstrated that – in multilinguals with comparable levels of proficiency – late L2 acquisition results in greater activation in L2 than in L1 (Vingerhoets et al., 2003; also shown by Perani et al., 2003; Wartenburger et al., 2003; Kovelman et al., 2008 in bilinguals). In quadrilingual subjects, although there was no clear association between the age of acquisition and the amount of activation, a negative correlation was found between the level of proficiency and the amount of activation (Briellmann et al., 2004). This was also suggested by the data of Abutalebi et al. (2013b) in trilinguals. Bloch et al. (2009) focused on the relation between the age of L2 acquisition and the variability of regional brain activation in Broca's and Wernicke's areas in subjects speaking at least three languages fluently and where the L3 had been learned after the age of 9 years. They demonstrated that variability in the representation of the three languages of the individual is related to the age of acquisition of L2, which indicates that early exposure to more than one language gives rise to a language processing network that can accommodate late learned languages.

The present study is based on the work of Bloch et al. (2009) and analyses structural MRI data of subjects fluent in at least three languages. The design of the study allowed us to search for structural differences between simultaneous and successive acquisition of L2. Thus, we suppose that simultaneous (SiM) versus successive or sequential acquisition of L1 and L2 (SuM) is associated with differences in the structural organization of brain areas subserving language processing. More precisely, we hypothesize that groups who acquired L2 later also show higher GM volumes in language-associated regions, as well as in other brain areas belonging to the extended language network (Ferstl et al., 2008). The design of the study provides us, further, with the opportunity to consider the potential role of the late L3 acquired by all participants. Structural differences between simultaneous (SiM) and successive multilinguals (SuMs) could indicate that very early acquired characteristics are maintained over a long period of life.

Materials and Methods

Subjects

Forty-four healthy, right-handed [verified by the outcome of the Edinburgh Handedness Inventory (Oldfield, 1971)], non-smoking multilinguals voluntarily participated in this study after receiving information about the investigation and the scanning process and giving their written informed consent. The subjects' average age at MRI acquisition was 28 years (range 18–37 years) and their female/male ratio was 22/22. The Ethics Committee of the University Hospital of Basel (EKBB, Switzerland) approved the study and confirmed its compliance with all relevant regulatory standards. All subjects were fluent and of medium to high proficiency in at least three languages (see **Table 1**). They did not differ with respect to their acquisition of their L3, which was comparable within all groups, and acquired at 9 years of age or later at school (see Bloch et al., 2009).

Assessment of Language Profiles

The multilinguals' age of second language acquisition was defined after analyzing each individual's language biography (Schütze, 1988; Schwabe, 2003) through oral interviews lasting 2–3 h. These

TABLE 1 | Language and proficiency profile of the 44 participants.

Group	Subgroup in Bloch et al. (2009)	L1		L2			L3	
			Level of competence		Level of competence	Level of immersion[1]		Level of competence
1 SiM	Simultaneous	English	*C2*	Swiss German	*C2*	High by context	Italian	*B2*
2 SiM	Simultaneous	Hungarian	*C2*	Swiss German	*C2*	High by context	English	*B1*
3 SiM	Simultaneous	Hungarian	*C1*	Swiss German	*C2*	High by context	English	*B1*
4 SiM	Simultaneous	Italian	*C1*	Swiss German	*B2+*	High by context	French	*B1+*
5 SiM	Simultaneous	French	*B2+*	Standard German	*C2*	High by context	English	*C2*
6 SiM	Simultaneous	French	*C2*	Standard German	*C1*	High by context	English	*B2+*
7 SiM	Simultaneous	Italian	*C1*	Standard German	*B2+*	High by context	French	*B1+*
8 SiM	Covert simultaneous	Italian	*C1*	Swiss German	*C2*	Medium-high by context	Spanish	*B1+*
9 SiM	Covert simultaneous	Italian	*C2*	Swiss German	*C2*	Medium-high by context	English	*B2*
10 SiM	Covert simultaneous	Greek	*B1+*	Swiss German	*C2*	Medium-high by context	Spanish	*C1*
11 SiM	Covert simultaneous	Slovene	*C2*	Swiss German	*C2*	Medium-high by context	English	*C1*
12 SiM	Covert simultaneous	French	*C2*	Standard German	*C2*	Medium-high by context	English	*B2*
13 SiM	Covert simultaneous	Serbo-Croatian	*C2*	Standard German	*C2*	Medium-high by context	English	*C1*
14 SiM	Covert simultaneous	Turkish	*C1*	Standard German	*C2*	Medium-high by context	English	*B2+*
15 SiM	Simultaneous	Standard German	*C2*	English	*B2*	High by family	French	*A2*
16 SiM	Simultaneous	Standard German	*C2*	Indonesian	*B1+*	High by family	English	*C1*
17 SiM	Simultaneous	Swiss German	*C2*	Italian	*C1*	High by family	English	*C1+*
18 SiM	Simultaneous	Swiss German	*C2*	Italian	*C2*	High by family	English	*C2*
19 SiM	Simultaneous	Spanish	*C2*	Catalan	*C2*	High by family	Swiss German	*C2*
20 SiM	Simultaneous	Spanish	*B2*	Catalan	*B1*	High by family	Standard German	*C2*
21 SiM	Simultaneous	Finish	*C2*	English	*C1*	High by family	Standard German	*C2*
22 SiM	Simultaneous	Portuguese	*B2*	French	*C1*		Japanese	*A1*
23 SiM[2]	Simultaneous	Catalan		Spanish			English	
24 SiM	Covert simultaneous	Bulgarian	*C2*	Russian	*C1*		French	*C1*
25 SuM	2nd to 5th year	Spanish	*C2*	Standard German	*C2*	High by context	English	*B1+*
26 SuM	2nd to 5th year	French	*B2*	Standard German	*C1*	High by context	English	*C2*
27 SuM	2nd to 5th year	Swiss German	*C2*	English	*C1*	Temporary high by context	French	*C2*
28 SuM	2nd to 5th year	Swiss German	*C2*	English	*C2*	Temporary high by context	Ivrit (New Hebrew)	*B2+*
29 SuM	2nd to 5th year	Swiss German	*C2*	English	*C1*	Temporary high by context	French	*B1*
30 SuM	2nd to 5th year	Standard German	*C2*	French	*C2*	Temporary high by context	English	*B1+*
31 SuM	2nd to 5th year	Standard German	*B2+*	French	*B2+*	Temporary high by context	English	*B1+*
32 SuM	2nd to 5th year	Spanish	*C1*	Italian	*B2*	Temporary high by context	Swiss German	*C2*
33 SuM	Late	Swiss German	*C2*	English	*C1*	Classroom learning	French	*C1*
34 SuM	Late	Standard German	*C2*	French	*B2+*	Classroom learning	Russian	*B2+*
35 SuM	Late	French	*C2*	Standard German	*B2+*	Classroom learning	English	*B1+*
36 SuM	Late	Swiss German	*C2*	English	*B2*	Classroom learning	French	*B2*
37 SuM	Late	Swiss German	*C2*	English	*C2*	Classroom learning	Italian	*B2*
38 SuM	Late	Swiss German	*C2*	English	*C2*	Classroom learning	French	*B2+*
39 SuM	Late	Italian	*C2*	Standard German	*B2*	Classroom learning	English	*B2*
40 SuM[2]	Late	Swiss German		French		Classroom learning	English	
41 SuM	Late	Swiss German	*C2*	French	*C2*	Classroom learning	English	*C2*
42 SuM	Late	Italian	*C2*	Standard German	*B2+*	Classroom learning	French	*B2*
43 SuM	Late	French	*C2*	English	*B2+*	Classroom learning	Standard German	*C2*
44 SuM	Late	Swiss German	*C2*	French	*B2*	Classroom learning	English	*B2*

Here, the classification of the multilingual participants and their languages with individual proficiencies levels is presented. Multilinguals are grouped into SiM and SuM, depending on their age of L2 acquisition. Prototypical details on level of immersion of L2 are given. A1 and A2 refer to competence levels of the basic user, B1 and B2 to the independent user, and C1 and C2 to the proficient user. A2+, B2+, and C1+ are intermediate stages [Common European Reference Framework for languages (CERR), Council of Europe, 2001; North, 2000]. Given the diglossic situation in the German speaking part of Switzerland, Standard German, and Swiss German can both be regarded as varieties of German. [1]*The classification into different levels of immersion is qualitative–descriptive and not quantitative-categorical and is aimed to give further input on the characteristics of our individual subjects rather than to mark or define strict groups of learners.* [2]*Subjects who did not return the CERR assessment form.*

in-depth linguistic biographies are based on the observation that free narrations give a more realistic account of the language history of the participant (Franceschini, 2002). The subjects could then be classified into four different groups of L2 acquisition: 16 ($F = 10$) simultaneous bilinguals, 8 ($F = 3$) covert simultaneous bilinguals, 8 ($F = 3$) sequential bilinguals, and 12 ($F = 6$) late multilinguals (Bloch et al., 2009). For the purpose of the present study, we re-grouped the participants into two groups:

(1) The *simultaneous multilingual* (SiM) group ($N = 24$; $F = 13$; average age: 27.7 years; age range: 18–36 years) consisting of simultaneous bilinguals, i.e., participants growing up in a bilingual family, where both parents/caregivers spoke different languages, and covert simultaneous bilinguals, i.e., participants growing up in a monolingual family whose language differed from that of the surroundings. This group of bilinguals was exposed to the L2 by the environment parallel to the L1, and for that reason they were grouped together with the simultaneous bilingual group.
(2) The *successive multilingual* (SuM) group ($N = 20$; $F = 9$; average age: 27.8 years; age range: 20–37 years) comprised successive bilinguals who had acquired their L2 subsequently to L1 between their second and fifth year of life, and late multilinguals who had acquired L2 at school when they were at least 9 years of age. Thus, this group covers multilinguals who acquired their L2 at a distinct time point *after* the acquisition of their L1.

Table 1 shows information on the multilingual profiles of the SiM and SuM groups. Additionally, it gives information about the type of L2 immersion. The level of competence in the individual languages was scaled by self-assessment, using the Common European Reference Framework for Languages (CERR, North, 2000; Council of Europe, 2001). Calculations using the Mann–Whitney *U* Test revealed no significant differences between SiM and SuM concerning L2 and L3 proficiency ($p > 0.05$, one-tailed).

Language Immersion Profile

As the age of acquisition is not the sole or exclusive influence to alter the structure of the brain, we provide here some additional, *post hoc* information on the *degree of immersion to L2* in our two groups of participants. Immersion has been shown to affect the brain's anatomy (Pliatsikas et al., 2015) and can be defined as the amount of naturalistic exposure, or immersion, that the speakers receive to that language. It is the degree to which language learners are exposed in their day-to-day activities, (see Pliatsikas and Chondrogianni, 2015).

This study was conducted in the German speaking part of Switzerland. This resulted in the recruited participants having significant exposure to German (the exceptions are subjects 20, 21, 22, 23, and 24 of the SiM group who either acquired Standard German in classroom circumstances as L3 after the age of 9 years or who did not report speaking German at all or learned it as an L4; and subjects 35, 39 and 42 of the SuM group who acquired Standard German in classroom circumstances as L2, see **Table 1**). Growing up in a German-speaking country makes the context for the acquisition of the L2 – in the cases when Standard German/Swiss German was learned as an L2 – one of early and *high immersion by context* (due to the linguistic dominance of the environment), or, in the case of the covert simultaneous-participants, a context of *medium–high immersion*. Similarly, growing up in a German speaking country with at least one caregiver speaking another language than German makes L2 acquisition of *high immersion by family*. Thus the majority of the members of SiM acquired their L2 by high immersion by family/context or medium-to-high immersion by context. The majority of the members of the SuM group learned L2 in a classroom setting. The subgroup of multilinguals classified as "2nd to 5th year" of age spent a period of their childhood/youth outside a German speaking country or moved to a German-speaking context. Their level of immersion to L2 is thus either *high by context* (2 out of 8) or *temporary high by context* (6 out of 8). Therefore, the quality of immersion differs between SiM and SuM (**Table 1**).

Magnetic Resonance Imaging

MR Image Acquisition

Magnetic resonance images of the 44 subjects were acquired on a 1.5-T Magnetom Vision MRI Scanner (Siemens, Erlangen, Germany) at the University Hospital of Basel. We used a standard head coil to restrict head movements and to limit motion artifacts. A three dimensional (3D) T_1-weighted anatomical high-resolution Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence was applied with repetition time of 9.7 ms, echo time of 4 ms, inversion time of 300 ms, and isotopic spatial resolution 1 mm × 1 mm × 1 mm (see also Bloch et al., 2009). The scans were all screened for major radiological abnormalities or visual artifacts by an experienced neuroradiologist.

Voxel-Based Morphometry

These MRI data were analyzed on commercially available Intel-based desktop computers with a Debian Linux 3.1 operating system. The structural images were pre-processed using a Voxel-Based Morphometry (VBM8) toolbox[1], as implemented in the Statistical Parametric Mapping software package[2]. The data were registered with "Diffeomorphic Anatomical Registration Through Exponentiated Lie" (DARTEL) within VBM8, running under the MATLAB 7.11.0 (R2010b) environment (Members of the Wellcome Trust Centre for Neuroimaging, 2009). Accuracy and sensitivity were maximized by creating a study-specific template and segmentation of each subject's image (Yassa and Stark, 2009). We conducted the following steps: (1) checking for scanner artifacts and major anatomical abnormalities for each subject; (2) aligning and reorientating the scans; (3) using New Segmentation and high-dimensional normalization DARTEL (Ashburner, 2007); (4) checking for homogeneity across the sample; and (5) using 8 mm standard smoothing (Bailey, 2008). The default values for realignment, warping, and normalization were used (Kurth et al., 2010). Finally, realigned, segmented, normalized, and smoothed data were subjected to statistical analysis.

[1] http://dbm.neuro.uni-jena.de/vbm8/
[2] http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

Of the 44 samples of this study two were outliers, with a mean covariance below 2 SDs. Repeated analyses without these two subjects did not change the results (not shown). We therefore decided to retain these two subjects.

Statistics

Voxel-based morphometry compares images on a voxel basis after spatial normalization using deformation fields that discount macroscopic differences in shape. We estimated between-group differences in GM volume at each intracerebral voxel in standard space by fitting a full-factorial analysis of covariance (ANCOVA), and contrasted the SiM and SuM groups. We modeled age at image acquisition and sex/gender as covariates of no interest, in order to reduce the potential impact of these variables on the GM volume in language-associated brain areas. We identified spatially continuous voxels at a threshold of $p < 0.01$ (uncorrected; cluster forming threshold; Petersson et al., 1999) and defined a family wise error-corrected cluster-extent threshold of $p < 0.05$ to infer statistical significance. In order to mark areas with significant GM volume differences on this statistical threshold level, Montreal Neurological Institute (MNI) coordinates were transformed into Talairach space (MNI and Talairach Transformation, 2013; Talairach.org Daemon, 2013).

Results

Gray matter volume was lower in the group of SiMs as compared to SuMs in the following regions: bilaterally in the medial frontal gyrus (MFG) and the left inferior frontal gyrus (IFG; $p < 0.001$, FWE); in the right IFG and right medial temporal gyrus (MTG); in the left inferior temporal gyrus (ITG); and in the right inferior posterior parietal gyrus ($p < 0.05$, FWE; **Table 2**; **Figure 1**). The opposite contrast SuM > SiM did not reveal any significant results.

Discussion

The present study of 44 multilinguals is to our knowledge the first VBM study in multilinguals as opposed to numerous studies

TABLE 2 | Results SiM < SuM.

ANCOVA full factorial 2 × 2, SiM < SuM. Height threshold $p_{uncorr} < 0.01$; extent threshold $k = 2497$ voxels.

	Area		MNI coordinates of cluster maximum (x/y/z)	Cluster $P_{FWE-corr}$	Cluster size k_E (voxels)	Peak level T
Frontal lobe	Medial frontal gyrus	L	−2/46/−9	<0.001	4107	4.71
		R	2/56/−11			
	Inferior frontal gyrus	L	−38/36/−8	<0.001	4753	4.61
		R	44/26/−2	0.015	2788	3.58
Temporal lobe	Inferior temporal gyrus	L	−63/−40/−24	0.016	2766	4.19
	Medial temporal gyrus	R	60/−27/−12	0.009	3071	3.90
Parietal lobe	Inferior posterior parietal gyrus	R	50/−79/48	0.027	2498	4.17

Clusters with significantly smaller gray matter volume in SiM are displayed.

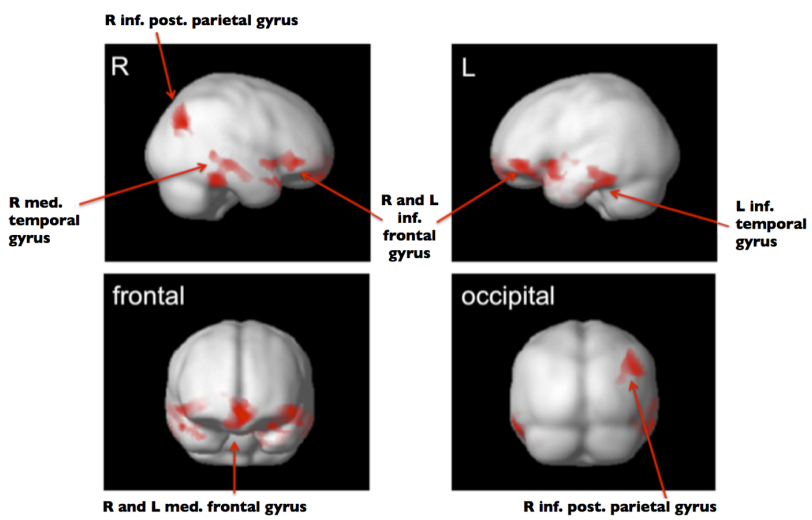


FIGURE 1 | ANCOVA full factorial 2 × 2. Significant clusters ($P_{FWE-corr} < 0.05$) with smaller GMV in SiM versus SuM shown in red. Group template used as background image in gray.

in bilinguals. It shows that GM volume was higher in the group of multilinguals who learned their L2 successively compared to the multilingual who acquired a second language simultaneously with their L1. Thus, subjects who did not acquire two languages simultaneously (by immersion) in early life but learned them sequentially, mostly in classroom settings, showed larger GM volume patterns in cortical language-associated regions and the extended language network (Ferstl et al., 2008). This result supports our thesis that early simultaneous bilingualism persists in the anatomical make-up of the adult brain.

The exact degree to which the difference between SiM and SuM is or is not influenced by a late learned L3 cannot be determined based on the present design. However, the other way round requires attention: testing trilinguals obliges us to tentatively hypothesize that despite the fact of a late learned L3 the differences based on the age of L2 acquisition persist into adulthood and do not disappear. Thus SiM and SuM remain different even though a late L3 was acquired. It can be assumed that the earlier in life a language experience is made, the more receptive the brain is to new learning and the more efficiently the brain can incorporate new language associated experiences, i.e., further input can be integrated into the same structural substrates. This would then be mirrored in GM patterns, particularly in the lower GM volume for early simultaneous L2 acquisition, including the late learned L3.

Our result corresponds with previous research showing an impact of L2 acquisition on GM volume of the bilingual brain (Mechelli et al., 2004; Osterhout et al., 2008; Stein et al., 2012) as well as research on GM volumetric changes in multilinguals (Elmer et al., 2014). However, the comparison of work applying GM density approaches and studies based on GM volumetric methods remains challenging. VBM is a technique that permits comparisons of the entire brain volume at the single voxel level. In contrast to previous studies (Mechelli et al., 2004; Osterhout et al., 2008; Stein et al., 2012) which reported GM densities rather than volumes, we used "optimized" VBM, which includes an additional modulation step to minimize the potentially confounding effects of errors in stereotactic normalization (Ashburner and Friston, 2001; Good et al., 2001). All images were smoothed using a 8 mm full-width-at-half-maximum Gaussian kernel, as in a previous study (Mechelli et al., 2004). According to the matched filter theorem, the width of the smoothing kernel determines the scale at which morphological changes are most sensitively detected (White et al., 2001). In the present study, we have chosen a rather small smoothing kernel, as this allows us to detect a greater number of regions with small structures and to better compare our results with the GM densities reported by Mechelli et al. (2004). Furthermore, both samples are based on healthy participants, so that we did not expect that regional changes would be very large or would differ much between cortical regions or between the studies. As expected, our results basically confirm the correlation (Mechelli et al., 2004; Stein et al., 2012) and/or association between age of L2 acquisition and GM structure, as reported previously (Osterhout et al., 2008).

In the present study, differences in GM volume were detected in the temporal as well as inferior and medial frontal regions of both hemispheres and the inferior parietal area. These are broadly parts of regions involved in the functional anatomy of language (Price, 2010) and may be linked, with the exception of the right inferior prefrontal cortex, to the "extended language network" outlined in influential work on the functional processing of language comprehension (Ferstl, 2007; Ferstl et al., 2008). Typically, primary language areas for language comprehension and production, such as Wernicke's and Broca's areas, did not display any significant difference between the SiM and SuM group. In the related fMRI study on multilingualism by Bloch et al. (2009), no specific group-dependent differences in activation were found in these areas either, which shows, for our population, that these regions are used, irrespectively of when a language is learned. The degree of variability in which they were activated by the three languages is, however, highly dependent on the age of L2 acquisition.

In the following, our GM data are discussed in relation to other studies on structural changes of gray matter linked to overall second language proficiency (Stein et al., 2014). Special attention is given to the bilateral character of our results. In their whole brain analysis of early and late bilinguals, Mechelli et al. (2004) found structural changes in the inferior parietal cortex (IPC) in relation to age; similarly Osterhout et al. (2008) employed an ROI analysis and detected structural alterations in the same region. Here too, GM changes in the IPC were detected and linked to age of acquisition. Others have demonstrated that the GM density of this region is positively correlated with vocabulary acquisition and knowledge, suggesting that this area is important not only for global L2 acquisition but for handling a large vocabulary (Lee et al., 2007; Richardson et al., 2010). Contrary to Mechelli et al. (2004), Osterhout et al. (2008), and Richardson et al. (2010), however, the present work showed changes in the right (x/y/z, 50/−79/48) and not in the left hemispheric IPC and thus supports the conclusions of Lee et al. (2007), who showed that vocabulary mastery predicted GM density in the bilateral IPC.

While Stein et al. (2012), in their longitudinal study on L2 acquisition, reported an increase in GM density in the left inferior frontal cortex (IFC) in close vicinity to the pars triangularis in Broca's area, the present study detected a bilateral pattern in this region and revealed a difference with respect to the age of L2 acquisition: Stein et al.'s (2012) participants learned L2 as adults, whereas our group of participants acquired their L2 as children. The results in our group of SiMs in IFC is in line with data of a recent cortical thickness study concerning the bilaterality of the structural patterns in this very same region (Klein et al., 2014). Klein et al. (2014) report, however, that thickness correlates positively with age of acquisition in the left IFG and negatively in the right IFG; this opposing interhemispheric effect was not re-enacted in our study where the association of age with higher GM volumes clearly counts for both the right and left IFG.

Our results on bilateral differences between SiM and SuM in GM volume in IFC are corroborated by functional data from bi- and multilinguals showing greater bilateral activation during both L1 and L2 processing than for monolinguals (Hull and Vaid, 2006; Park et al., 2012), with a strong tendency for the right Broca's homolog to be activated in the L1 of multilinguals (Kaiser et al., 2007). There is still little evidence about the function of the

MFG in the context of structural changes due to L2 acquisition, except that its cortical thickness shows alterations after very intensive language training over many months (Mårtensson et al., 2012). Again, our present results reveal structural bilateral modifications in the MFG, whereas Mårtensson et al. (2012) found these only in the left side. Data based on functional MRI studies suggest that the MFG is crucial for text comprehension (Ferstl and von Cramon, 2001).

It is well known that both the inferior and the middle temporal gyrus handle various aspects of lexical semantic representation and processing. Our study is, to our knowledge, the first to present changes in the GM volume due to early L2 acquisition in both the right MTG and left ITG.

Taken together, the right hemispheric trend (as exemplified in r-IPC; bilateral IFC; bilateral MFG; r-MTG) – which characterizes our set of multilinguals when differentiated by the age of L2 acquisition – could be influenced by two additional factors: by the mode of L2 acquisition and/or by the interference of the L3.

Our subjects were carefully selected on the basis of their multilingual profile and had to undergo extensive interviewing for 2–3 h about their three languages. However, some information about their languages was not captured. Thus, for instance, it cannot be excluded with certainty whether any of the simultaneous participants registered with German/Swiss German as L2 grew up in a non-German speaking country and acquired L1 from one caregiver and L2 (German/Swiss German) from a second caregiver. Nevertheless, the presence of immersion as the main access to L2 acquisition remains valid in SiM, as well as the presence of classroom learning in SuM. Future cross-sectional and longitudinal research is needed to identify which of these ways of learning L2, i.e., L2 acquisition based on *high immersion by family*, *high immersion by context*, *medium-high immersion by context, temporary high immersion by context*, or *classroom learning* has a greater impact on structural GM in relevant brain areas. Most recently, Pliatsikas et al. (2015) demonstrated the effects of immersion on brain structure in young, highly immersed late bilinguals. In their view, "[immersion] can be broadly defined as the degree to which language learners use their non-native language outside the classroom and for their day-to-day activities and usually presupposes that the learners live in an environment where their non-native language is exclusively or mostly used" (see Pliatsikas and Chondrogianni, 2015). Interestingly, structural alterations in white matter were shown to be effected by everyday L2 use in a naturalistic environment, rather than by length of L2 learning or age of onset of L2 learning (Pliatsikas et al., 2015). Thus, our results in the group of SiM can also be interpreted from this perspective showing that naturalistic exposure, rather than age of L2 acquisition, impacts on brain structure. This is not the case for the group of SuM who in their majority acquired L2 as a foreign language in the classroom.

Finally, the impact of the L3 on language-associated brain areas and the extended language network remains to be elucidated. In our experimental setup, the age of L2 acquisition is the variable determining the GM structure early in development. However, the way in which training for additional languages drives GM plasticity in regions already influenced by bilingualism is open to speculation. Additional training might result in pruning of language networks, as suggested by Elmer et al. (2014), or might drive contralateral (right hemisphere) cortical areas to participate in language related tasks.

The results of the present study cannot prove a lifelong plasticity of the brain for languages since the examined subjects were chosen based on the fact of having learnt at least three languages during early or late childhood, respectively, and we do not know about further changes in their brains during adulthood. Neither do the obtained outcomes give any information if there is a critical age for the native-like acquisition of one or many languages, although there are certainly studies showing complex relations between the maturation of the brain in children and the brain's plasticity to adjust to structural demands of (individual) language development (Brauer et al., 2011).

Methodical Considerations

Modification of the user-options implemented in the analysis software, such as setting the smoothing kernel, can influence subsequent statistical results (Ashburner and Friston, 2001): a larger kernel (12 mm instead of 8 mm) results in greater cluster sizes (Friston, 2003, p. 5). We used a default option of 8 mm as recommended. VBM compares voxel-by-voxel images of different groups and reports MNI standard coordinates for every cluster center, which does not necessarily correspond to the actual cluster localisation in an individual's brain. Thus VBM statistics do not differentiate between two clusters localized for example in the medial plane. In the present case, the large clusters in the medial frontal gyri on the left and the right are counted as one cluster.

As to the statistical thresholds, it is still very difficult to compare data from different studies. Height thresholds range from $p_{\text{uncorr}} < 0.001$ up to $p_{\text{corr}} < 0.05$ in different studies (Brambati et al., 2004; Eckert et al., 2005; Silani et al., 2005; Hoeft et al., 2007; Steinbrink et al., 2008; Richardson and Price, 2009). Here, an uncorrected height threshold of $p < 0.01$ was used. The clusters found are therefore large, although there are small differences between the two groups.

Conclusion

Contrary to the successive acquisition of the second language, simultaneous acquisition of L1/L2 (by immersion) from the first year of life on is associated with low GM volume in language-associated regions, in the prefrontal, medial temporal and parietal cortex, in particular. This difference persists even though a late L3 is learned. Growing up in a multilingual environment in early childhood may change the individual's cortical structure, enforcing it to generally build more efficient synaptic networks for language processing. To further understand structural changes underlying brain plasticity during language learning requires longitudinal studies with homogenous groups of SiM and SuM.

References

Abutalebi, J., Della Rosa, P. A., Castro Gonzaga, A. K., Keim, R., Costa, A., and Perani, D. (2013a). The role of the left putamen in multilingual language production. *Brain Lang.* 125, 307–315. doi: 10.1016/j.bandl.2012.03.009

Abutalebi, J., Della Rosa, P., Ding, G., Weekes, B., Costa, A., and Green, D. (2013b). Language proficiency modulates the engagement of cognitive control areas in multilinguals. *Cortex* 49, 905–911. doi: 10.1016/j.cortex.2012.08.018

Abutalebi, J., Della Rosa, P. A., Green, D. W., Hernandez, M., Scifo, P., Keim, R., et al. (2012). Bilingualism tunes the anterior cingulate cortex for conflict monitoring. *Cereb. Cortex* 22, 2076–2086. doi: 10.1093/cercor/bhr287

Amunts, K., Schleicher, A., and Zilles, K. (2004). Outstanding language competence and cytoarchitecture in Broca's speech region. *Brain Lang.* 89, 346–353. doi: 10.1016/S0093-934X(03)00360-2

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113.

Ashburner, J., and Friston, K. (2001). Why voxel-based morphometry should be used. *Neuroimage* 14, 1238–1243. doi: 10.1006/nimg.2001.0961

Bailey, R. A. (2008). *Design of Comparative Experiments*. Cambridge: Cambridge University Press.

Bloch, C., Kaiser, A., Kuenzli, E., Zappatore, D., Haller, S., Franceschini, R., et al. (2009). The age of second language acquisition determines the variability in activation elicited by narration in three languages in Broca's and Wernicke's area. *Neuropsychologia* 47, 625–633. doi: 10.1016/j.neuropsychologia.2008.11.009

Brambati, S. M., Termine, C., Ruffino, M., Stella, G., Fazio, F., Cappa, S. F., et al. (2004). Regional reductions of grey matter volume in familial dyslexia. *Neurology* 63, 742–745.

Brauer, J., Anwander, A., and Friederici, A. D. (2011). Neuroanatomical prerequisites for language functions in the maturing brain. *Cereb. Cortex* 21, 459–466. doi: 10.1093/cercor/bhq108

Briellmann, R. S., Saling, M. M., Connell, A. B., Waites, A. B., Abbott, D. F., and Jackson, G. D. (2004). A high-field functional MRI study of quadri-lingual subjects. *Brain Lang.* 89, 531–542. doi: 10.1016/j.bandl.2004.01.008

Coggins, P. E. III, Kennedy, T. J., and Armstrong, T. A. (2004). Bilingual corpus callosum variability. *Brain Lang.* 89, 69–75. doi: 10.1016/S0093-934X(03)00299-2

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J. T., et al. (2006). Language control in the bilingual brain. *Science* 312, 1537–1540. doi: 10.1126/science.1127761

Cummine, J., and Boliek, C. A. (2013). Understanding white matter integrity stability for bilinguals on language status and reading performance. *Brain Struct. Funct.* 218, 595–601. doi: 10.1007/s00429-012-0466-6

Eckert, M. A., Leonard, C. M., Wilke, M., Eckert, M., Richards, T., Richards, A., et al. (2005). Anatomical signatures of dyslexia in children: unique information from manual and voxel based morphometry brain measures. *Cortex* 41, 304–315.

Elmer, S., Hänggi, J., and Jäncke, L. (2014). Processing demands upon cognitive, linguistic, and articulatory functions promote grey matter plasticity in the adult multilingual brain: insights from simultaneous interpreters. *Cortex* 54, 179–189. doi: 10.1016/j.cortex.2014.02.014

Elmer, S., Hänggi, J., Meyer, M., and Jäncke, 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. doi: 10.1002/hbm.21169

Ferstl, E. (2007). "The functional neuroanatomy of text comprehension: what's the story so far?" in *Higher Level Language Processes in the Brain: Inference and Comprehension Processes*, eds F. Schmalhofer and C. Perfetti (Mahwah, NJ: Lawrence Erlbaum), 53–102.

Ferstl, E., Neumann, J., Bogler, C., and von Cramon, D. Y. (2008). The extended language network: a meta-analysis of neuroimaging studies on text comprehension. *Hum. Brain Mapp.* 29, 581–593. doi: 10.1002/hbm.20422

Ferstl, E., and von Cramon, D. (2001). The role of coherence and cohesion in text comprehension: an event-related fMRI study. *Cogn. Brain Res.* 11, 325–340. doi: 10.1016/S0926-6410(01)00007-6

Filippi, R., Richardson, F. M., Dick, F., Leech, R., Green, D. W., Thomas, M. S. C., et al. (2011). The right posterior paravermis and the control of language interference. *J. Neurosci.* 31, 10732–10740. doi: 10.1523/JNEUROSCI.1783-11.2011

Franceschini, R. (2002). "Sprachbiographien: Erzählungen über Mehrsprachigkeit und deren Erkenntnisinteresse für die Spracherwerbsforschung und die Neurobiologie der Mehrsprachigkeit," in *Biografie linguistiche/Biographies langagières/Biografias linguisticas/Sprachbiografien*, eds K. Adamzik and E. Roos (Neuchâtel: Bulletin vals-asla), 19–33.

Friston, K. J. (2003). Statistical parametric mapping. *Neurosci. Databases* 4, 237–246. doi: 10.1006/nimg.2002.1259

García-Pentón, L., Fernández, A. P., Iturria-Medina, Y., Gillon-Dowens, M., and Carreiras, M. (2014). Anatomical connectivity changes in the bilingual brain. *Neuroimage* 84, 495–504. doi: 10.1016/j.neuroimage.2013.08.064

Golestani, N., Molko, N., Dehaene, S., LeBihan, D., and Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cereb. Cortex* 17, 575–582. doi: 10.1093/cercor/bhk001

Golestani, N., Paus, T., and Zatorre, R. J. (2002). Anatomical correlates of learning novel speech sounds. *Neuron* 35, 997–1010.

Golestani, N., Price, C. J., and Scott, S. K. (2011). Born with an ear for dialects? Structural plasticity in the expert phonetician brain. *J. Neurosci.* 31, 4213–4220. doi: 10.1523/JNEUROSCI.3891-10.2011

Good, C., Johnsrude, I., Ashburner, J., Henson, R., Friston, K., and Frackowiak, R. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* 14, 21–36. doi: 10.1006/nimg.2001.0786

Grogan, A., Green, D. W., Ali, N., Crinion, J. T., and Price, C. J. (2009). Structural correlates of semantic and phonemic fluency ability in first and second languages. *Cereb. Cortex* 19, 2690–2698. doi: 10.1093/cercor/bhp023

Grogan, A., Parker Jones, O., Ali, N., Crinion, J., Orabona, S., Mechias, M. L., et al. (2012). Structural correlates for lexical efficiency and number of languages in non-native speakers of English. *Neuropsychologia* 50, 1347–1352. doi: 10.1016/j.neuropsychologia.2012.02.019

Hoeft, F., Meyler, A., Hernandez, A., Juel, C., Taylor-Hill, J., Martindale, J., et al. (2007). Functional and morphometric brain dissociation between dyslexia and reading ability. *Proc. Natl. Acad. Sci. U.S.A.* 104, 4234–4239. doi: 10.1073/pnas.0609399104

Hosoda, C., Tanaka, K., Nariai, T., Honda, M., and Hanakawa, T. (2013). Dynamic neural network reorganization associated with second language vocabulary acquisition: a multimodal imaging study. *J. Neurosci.* 33, 13663–13672. doi: 10.1523/JNEUROSCI.0410-13.2013

Hull, R., and Vaid, J. (2006). Laterality and language experience. *Laterality* 11, 436–464. doi: 10.1080/13576500600691162

Kaiser, A., Kuenzli, E., Zappatore, D., and Nitsch, C. (2007). On females' lateral and males' bilateral activation during language production: a fMRI study. *Int. J. Psychophysiol.* 63, 192–198. doi: 10.1016/j.ijpsycho.2006.03.008

Klein, D., Mok, K., Chen, J. K., and Watkins, K. E. (2014). Age of language learning shapes brain structure: a cortical thickness study of bilingual and monolingual individuals. *Brain Lang.* 131, 20–24. doi: 10.1016/j.bandl.2013.05.014

Kovelman, I., Baker, S. A., and Petitto, L.-A. (2008). Bilingual and monolingual brains compared: a functional magnetic resonance imaging investigation of syntactic processing and a possible 'neural signature' of bilingualism. *J. Cogn. Neurosci.* 20, 153–169. doi: 10.1162/jocn.2008.20011

Kurth, F., Luders, E., and Gaser, C. (2010). *VBM8-Toolbox Manual*. Jena: University of Jena.

Lee, H. L., Devlin, J. T., Shakeshaft, C., Stewart, L. H., Brennan, A., Glensman, J., et al. (2007). Anatomical traces of vocabulary acquisition in the adolescent brain. *J. Neurosci.* 27, 1184–1189. doi: 10.1523/JNEUROSCI.4442-06.2007

Li, P., Legault, J., and Litcofsky, K. A. (2014). Neuroplasticity as a function of second language learning: anatomical changes in the humanbrain. *Cortex* 58, 301–324. doi: 10.1016/j.cortex.2014.05.001

Mårtensson, J., Eriksson, J., Bodammer, N. C., Lindgren, M., Johansson, M., Nyberg, L., et al. (2012). Growth of language-related brain areas

after foreign language learning. *Neuroimage* 63, 240–244. doi: 10.1016/j.neuroimage.2012.06.043

Mechelli, A., Crinion, J. T., Noppeney, U., O'Doherty, J., Ashburner, J., Frackowiak, R. S., et al. (2004). Neurolinguistics: structural plasticity in the bilingual brain. *Nature* 431, 757. doi: 10.1038/431757a

Members of the Wellcome Trust Centre for Neuroimaging. (2009). *SPM8 Statistical Parametric Mapping (Version 8)*. Available at: http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

MNI and Talairach Transformation. (2013). Available at: http://www.ebire.org/hcnlab/cortical-mapping/ (accessed March 17, 2013).

Mohades, S. G., Struys, E., Van Schuerbeek, P., Mondt, K., Van De Craen, P., and Luypaert, R. (2012). DTI reveals structural differences in white matter tracts between bilingual and monolingual children. *Brain Res.* 1435, 72–80. doi: 10.1016/j.brainres.2011.12.005

Nauchi, A., and Sakai, K. L. (2009). Greater leftward lateralization of the inferior frontal gyrus in second language learners with higher syntactic abilities. *Hum. Brain Mapp.* 30, 3625–3635. doi: 10.1002/hbm.20790

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. Ph.D. thesis, Thames Valley University, New York, NY: Peter Lang.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113.

Osterhout, L., Poliakov, A., Inoue, K., McLaughlin, J., Valentine, G., Pitkanen, I., et al. (2008). Second-language learning and changes in the brain. *J. Neurolinguistics* 21, 509–521. doi: 10.1016/j.jneuroling.2008.01.001

Park, H. R. P., Badzakova-Trajkov, G., and Waldie, K. E. (2012). Language lateralisation in late proficient bilinguals: a lexical decision fMRI study. *Neuropsychologia* 50, 688–695. doi: 10.1016/j.neuropsychologia.2012.01.005

Perani, D., Abutalebi, J., Paulesu, E., Brambati, S., Scifo, P., Cappa, S. F., et al. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: an fMRI study during verbal fluency. *Hum. Brain Mapp.* 19, 170–182. doi: 10.1002/hbm.10110

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. (1999). Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 1261–1281.

Pliatsikas, C., and Chondrogianni, V. (2015). Learning a non-native language in a naturalistic environment: insights from behavioral and neuroimaging research (Editorial note). Frontiers Special Issue.

Pliatsikas, C., Johnstone, T., and Marinis, T. (2014). Grey matter volume in the cerebellum is related to the processing of grammatical rules in a second language: a structural voxel-based morphometry study. *Cerebellum* 13, 55–63. doi: 10.1007/s12311-013-0515-6

Pliatsikas, C., Moschopoulou, E., and Saddy, J. D. (2015). The effects of bilingualism on the white matter structure of the brain. *Proc. Nat. Acad. Sci. U.S.A.* 112, 1334–1337. doi: 10.1073/pnas.1414183112

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* 1191, 62–88. doi: 10.1111/j.1749-6632.2010.05444.x

Ressel, V., Pallier, C., Ventura-Campos, N., Díaz, B., Roessler, A., Ávila, C., et al. (2012). An effect of bilingualism on the auditory cortex. *J. Neurosci.* 32, 16597–16601. doi: 10.1523/JNEUROSCI.1996-12.2012

Richardson, F., Michael, T., Filippi, R., Harth, H., and Price, C. J. (2010). Contrasting effects of vocabulary knowledge on temporal and parietal brain structure across lifespan. *J. Cogn. Neurosci.* 22, 943–954. doi: 10.1162/jocn.2009.21238

Richardson, F. M., and Price, C. J. (2009). Structural MRI studies of language function in the undamaged brain. *Brain Struct. Funct.* 213, 511–523. doi: 10.1007/s00429-009-0211-y

Schlegel, A. A., Rudelson, J. J., and Tse, P. U. (2012). White matter structure changes as adults learn a second language. *J. Cogn. Neurosci.* 24, 1664–1670. doi: 10.1162/jocn_a_00240

Schütze, F. (1988). *Das Narrative Interview in Interaktionsfeldstudien*. Hagen: Fernuniversität Gesamthochschule Hagen.

Schwabe, M. (2003). Review: Gabriele Lucius-Hoene & Arnulf Deppermann (2002). Rekonstruktion narrativer Identität. Ein Arbeitsbuch zur Analyse narrativer Interviews [The Reconstruction of Narrative Identity. A Guide to the Analysis of Narrative Interviews]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 4. Avaialble at: http://www.qualitative-research.net/index.php/fqs/article/view/691/1492

Silani, G., Frith, U., Demonet, J. F., Fazio, F., Perani, D., Price, C., et al. (2005). Brain abnormalities underlying altered activation in dyslexia: a voxel based morphometry study. *Brain* 128, 2453–2461. doi: 10.1093/brain/awh579

Stein, M., Federspiel, A., Koenig, T., Wirth, M., Strik, W., Wiest, R., et al. (2012). Structural plasticity in the language system related to increased second language proficiency. *Cortex* 48, 458–465. doi: 10.1016/j.cortex.2010.10.007

Stein, M., Winkler, C., Kaiser, A., and Dierks, T. (2014). Structural brain changes related to bilingualism: does immersion make a difference? *Front. Psychol.* 5:1116. doi: 10.3389/fpsyg.2014.01116

Steinbrink, C., Vogt, K., Kastrup, A., Müller, H.-P., Juengling, F. D., Kassubek, J., et al. (2008). The contribution of white and grey matter differences to developmental dyslexia: insights from DTI and VBM at 3.0 T. *Neuropsychologia* 46, 3170–3178. doi: 10.1016/j.neuropsychologia.2008.07.015

Talairach.org Daemon. (2013). Available at: http://www.talairach.org/daemon.html (accessed March 17, 2013).

Taubert, M., Draganski, B., Anwander, K., Müller, K., Horstmann, A., Villringer, A., et al. (2010). Dynamic properties of human brain structure: learning-related changes in cortical areas and associated fiber connections. *J. Neurosci.* 30, 11670–11677. doi: 10.1523/JNEUROSCI.2567-10.2010

Vingerhoets, G., Van Borsel, J., Tesink, C., van den Noort, M., Deblaere, K., Seurinck, R., et al. (2003). Multilingualism: an fMRI study. *Neuroimage* 20, 2181–2196.

Wartenburger, I., Heekeren, H. R., Abutalebi, J., Cappa, S. F., Villringer, A., and Perani, D. (2003). Early setting of grammatical processing in the bilingual brain. *Neuron* 37, 159–170.

White, T., O'Leary, D., Magnotta, V., Arndt, S., Flaum, M., and Andreasen, N. (2001). Anatomic and functional variability: the effects of filter size in group fMRI data analysis. *Neuroimage* 13, 577–588. doi: 10.1006/nimg.2000.0716

Wong, P. C. M., Warrier, C. M., Penhune, V. B., Roy, A. K., Sadehh, A., Parrish, T. B., et al. (2008). Volume of left Heschl's gyrus and linguistic pitch learning. *Cereb. Cortex* 18, 828–836. doi: 10.1093/cercor/bhm115

Yassa, M. A., and Stark, C. E. L. (2009). A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *Neuroimage* 44, 319–327. doi: 10.1016/j.neuroimage.2008.09.016

Zatorre, R. J., Fields, R. D., and Johansen-Berg, H. (2012). Plasticity in grey and white: neuroimaging changes in brain structure during learning. *Nat. Neurosci.* 15, 528–536. doi: 10.1038/nn.3045

Zhang, M., Li, J., Chen, C., Mei, L., Xue, G., Lu, Z., et al. (2013). The contribution of the left mid-fusiform cortical thickness to Chinese and English reading in a large Chinese sample. *Neuroimage* 65, 250–256. doi: 10.1016/j.neuroimage.2012.09.045

Zou, L., Abutalebi, J., Zinszer, B., Yan, X., Shu, H., Peng, D., et al. (2012). Second language experience modulates functional brain network for the native language production in bimodal bilinguals. *NeuroImage* 62, 1367–1375. doi: 10.1016/j.neuroimage.2012.05.062

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.