WHAT ARE (UN)ACCEPTABILITY AND (UN)GRAMMATICALITY? HOW DO THEY RELATE TO ONE ANOTHER AND TO INTERPRETATION?

EDITED BY: Susagna Tubau, Urtzi Etxeberria, Viviane Marie Deprez and M.Teresa Espinal

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-374-3
DOI 10.3389/978-2-88966-374-3

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

WHAT ARE (UN)ACCEPTABILITY AND (UN)GRAMMATICALITY? HOW DO THEY RELATE TO ONE ANOTHER AND TO INTERPRETATION?

Topic Editors:
Susagna Tubau, Autonomous University of Barcelona, Spain
Urtzi Etxeberria, Centre National de la Recherche Scientifique (CNRS), France
Viviane Marie Deprez, Centre National de la Recherche Scientifique (CNRS), France
M.Teresa Espinal, Autonomous University of Barcelona, Spain

Citation: Tubau, S., Etxeberria, U., Deprez, V. M., Espinal, M. T., eds. (2021). What are (Un)Acceptability and (Un)Grammaticality? How do They Relate to One Another and to Interpretation?. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-374-3

Table of Contents

frontiers in Psychology

EDITORIAL
published: 03 December 2020
doi: 10.3389/fpsyg.2020.621267

Editorial: What Are (Un)Acceptability and (Un)Grammaticality? How Do They Relate to One Another and to Interpretation?

Susagna Tubau[1]*, *Urtzi Etxeberria*[2], *Viviane Déprez*[3,4] and *M.Teresa Espinal*[5]

[1] Departament de Filologia Anglesa i de Germanística, Universitat Autònoma de Barcelona, Barcelona, Spain, [2] Centre National de la Recherche Scientifique, IKER (UMR5478), Bayonne, France, [3] Department of Linguistics, Rutgers University, New Brunswick, NJ, United States, [4] Centre National de la Recherche Scientifique, ISC Marc Jeannerod (UMR5304), Bron, France, [5] Departament de Filologia Catalana, Universitat Autònoma de Barcelona, Barcelona, Spain

Keywords: (un)acceptability, (un)grammaticality, interpretation, linguistics, experimental investigation

Editorial on the Research Topic

What Are (Un)Acceptability and (Un)Grammaticality? How Do They Relate to One Another and to Interpretation?

That grammatical sentences and their interpretation form the building blocks of linguistic theories is not controversial. Yet, the collection of articles in the present Research Topic shows that the notion of (un)grammaticality, on the one hand, and the observations of (un)acceptability ratings, on the other, can entertain in fact rather complex interactions. That is, the relation between the notion of grammaticality and the actual acceptability that speakers attribute to sentences is far from being straightforward: not only can some grammatical sentences present parsing difficulties that cause speakers to judge them unacceptable, but also sentences that are considered ungrammatical by linguists could be perceived as acceptable by speakers and lead to reliable interpretations. In addition, the methodology used in the investigation of (un)acceptability and (un)grammaticality and their relation may play an important role in our ultimate understanding of these two core notions which, despite being in principle independent from one another, often crisscross. Therefore, it seems useful and perhaps necessary to engage in actively evaluating how certain research methods can prove particularly useful when trying to establish the degree and extent to which (un)grammatical linguistic structures and their interpretations are (un)acceptable to speakers, and how this can be taken to reliably and consistently relate to (un)grammaticality.

As discussed in the Hypothesis and Theory article by Leivada and Westergaard, the relation between grammaticality, acceptability (and parsability), as found in the literature, is in need of terminological clarification, as (un)grammaticality and (un)acceptability do not homogeneously manifest coincident scales. Actually, further empirical confirmation that ungrammaticality can correspond to a speaker's misperception is found in the Original Research article by de-Dios-Flores where so called negative polarity item illusions in English are investigated. The author shows that grammatical sentences with multiple negations can be perceived as unacceptable under certain processing conditions. This complements previous research showing that ungrammatical sentences could be perceived as acceptable.

In a similar vein, the Original Research article by Blanchette and Lukyanenko offers empirical support to the idea that acceptability and grammaticality are not necessarily equated, thus making it possible for the grammar of English speakers to generate Negative Concord structures that are nonetheless judged with low acceptability ratings due to contextual factors. The Original Research article by Hubers et al. also illustrates that speakers of a language can use linguistic constructions

Edited and reviewed by:
Antonio Benítez-Burraco, Sevilla University, Spain

***Correspondence:**
Susagna Tubau
Susagna.Tubau@uab.cat

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

Received: *25 October 2020*
Accepted: *02 November 2020*
Published: *03 December 2020*

Citation:
Tubau S, Etxeberria U, Dëprez V and Espinal MT (2020) Editorial: What Are (Un)Acceptability and (Un)Grammaticality? How Do They Relate to One Another and to Interpretation? Front. Psychol. 11:621267. doi: 10.3389/fpsyg.2020.621267

that violate prescriptive rules. In their article they empirically investigate how one particular instance of these grammatical norm violations (i.e., the use of the equative particle instead of the comparative particle) is processed in Dutch and German. Results of three different experiments show that they are processed differently both from ungrammatical and grammatical sentences.

Also closely connected to acceptability is the frequency of occurrence in the context of language variation, as shown in the Original Research article by Gerasimova and Lyutikova. The authors address how different grammatical variants available to a single speaker in Russian distribute in production and perception, the main finding being that the more frequent a variant is, the higher the acceptability score speakers attribute to it. Nevertheless, the variants that are perceived as highly acceptable by the speakers are not always the ones that occur more frequently in production.

That methodological issues are relevant to the definition of acceptability and grammaticality is shown in the Original Research article by Langsford et al., who manipulate the instructions given to the participants of an experiment consisting in evaluating the acceptability/grammaticality of stimuli sentences to investigate whether instructions can help control variability in the motivation underlying ratings that has been identified in the literature. Their results show that participants indeed rate the sentences differently depending on whether they are asked to consider their acceptability or their grammaticality (the latter judgements being more extreme than the former).

The Brief Research Report by Gavarró shows that it is possible to use grammaticality judgement tasks with children, predicting that the differences in production and comprehension that children may display in comparison to adults will also show in a grammaticality judgement task, as production, comprehension, and grammaticality judgements would align. The author uses this methodology to investigate Relativized Minimality in child Catalan and argues that it can help determine whether it constitutes a grammatical or a processing phenomenon. The grammaticality judgement task is also the methodology used in Perpiñán's Original Research article on the sensitivity of L1 and L2 speakers of Spanish to extraction from island configurations. Perpiñán's experimental results show that L2 learners and native speakers use the same processing and interpretative mechanisms for parsing islands and point at the need to redefine grammaticality more holistically, as factors such as plausibility and processability might have a strong influence on it.

Oseki and Marantz use an acceptability judgement experiment to investigate morphologically complex words and evaluate five different computational models of morphological competence. Their results show that models with morpheme units outperform models without them. On the basis of the computational modeling of acceptability data, the authors show that morphological competence is best characterized as involving grammar-generated hierarchical structures rather than external surface linear strings in corpora.

On a related note, Huang and Ferreira discuss acceptability judgements in a Methods article. Acceptability judgements have been widely used in linguistic research, but have proven controversial, as they have a number of limitations and can include bias in the speakers' responses. This leads the authors to propose the application of Signal Detection Theory—a method used in other psychological research areas—to judgement data, with the aim of more effectively controlling bias. Further support that acceptability judgement methodology can blur conclusions on (un)acceptability is found in Wellwood's Hypothesis and Theory article. By considering a case study in degree semantics (i.e., adjective scale structure), the author proposes a two-step model of semantic interpretation that separates meaning from interpretation, and that relates language to thought.

In short, the articles in the present Research Topic confirm that it is indeed necessary to try to theoretically and empirically explore and (re)define (un)acceptability and (un)grammaticality as core notions that interact in complex ways, not only with one another, but also with the interpretation of sentences. This can certainly result not only in a better understanding of what makes (un)grammatical sentences (un)acceptable, but also of the role of performance factors, memory limitations, and processing mechanisms in the evaluation of (un)acceptability, (un)grammaticality, and the interpretation of linguistic structures. The methodological choices made when researching linguistic phenomena related to (un)acceptability, (un)grammaticality and/or their interaction have also been discussed as an essential piece of the research plan that should not be overlooked.

AUTHOR CONTRIBUTIONS

ST drafted the work, which was revised critically by UE, VD, and MTE. All authors contributed to the final approval of the work for publication.

ACKNOWLEDGMENTS

The editorial work in this volume has been supported by grants awarded by the Spanish Ministerio de Economía y Competitividad (FFI2017-82547-P), the Generalitat de Catalunya (2017SGR634), the Netherlands Organisation Committee (Partitivity in European Languages, PARTE), the ANR (BIM ANR-17-CE27-11), and the Franco-German ANR-DFG (UV2 ANR-18-FRAL-0006). MTE also acknowledges an Academia award from the Catalan Institution for Research and Advanced Studies (ICREA).

Conflict of Interest: 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
published: 22 October 2019
doi: 10.3389/fpsyg.2019.02346

Processing Sentences With Multiple Negations: Grammatical Structures That Are Perceived as Unacceptable

Iria de-Dios-Flores[1,2]*

[1] English and German Department, Universidade de Santiago de Compostela, Santiago de Compostela, Spain,
[2] Basque Center on Cognition, Brain and Language, Donostia, Spain

This investigation draws from research on negative polarity item (NPI) illusions in order to explore a new and interesting instance of misalignment observed for grammatical sentences containing two negative markers. Previous research has shown that unlicensed NPIs can be perceived as acceptable when occurring soon after a structurally inaccessible negation (e.g., *ever* in **The bills that no senators voted for have ever become law*). Here we examine the opposite configuration: grammatical sentences created by substituting the NPI *ever* with the negative adverb *never* (e.g., *The bills that no senators voted for have never become law*). The processing and acceptability of these sentences were studied using three tasks: a speeded acceptability judgment (Experiment 1), a self-paced reading task (Experiment 2), and an offline acceptability rating (Experiment 3). The results are consistent across measures in showing that the integration of the adverb *never* is disrupted by the linearly preceding but structurally inaccessible negative quantifier *no* in the relative clause. In our view, this pattern of results is in line with Parker and Phillips' (2016) proposal that NPI illusions arise when the context containing the inaccessible negation has not been fully encoded by the time the NPI *ever* is encountered, making the embedded negative quantifier transparently available as a licensor. In a similar vein, the disruption effects observed for grammatical sentences containing two negative elements could arise if the negative quantifier is still being integrated when *never* is encountered, forcing the parser to deal with two negative elements simultaneously. This interpretation suggests that the same incomplete encodings that could be ameliorating the online perception of unlicensed NPIs could also be responsible for deteriorating the perception of the sentences under investigation here. This would represent an illusion of ungrammaticality. Furthermore, these results provide evidence against the speculation that NPI illusions are the consequence of misrepresenting *ever* as its near neighbor *never*, given that continuations with *never* are judged as unacceptable in spite of their grammaticality. Together, these findings inform the landscape of hypotheses on NPI illusions and offer valuable insights into the complexity of multiple negations and the relation between processing difficulty and acceptability.

Keywords: multiple negation, double negation, acceptability, grammatical illusions, interference, negative polarity items, processing complexity

OPEN ACCESS

Edited by:
Susagna Tubau, Autonomous University of Barcelona, Spain

Reviewed by:
Jeremy Pasquereau, University of Surrey, United Kingdom
Brian Dillon, University of Massachusetts Amherst, United States

***Correspondence:**
Iria de-Dios-Flores
iria.dedios@usc.gal;
iriadediosflores@gmail.com

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

Received: *10 June 2019*
Accepted: *01 October 2019*
Published: *22 October 2019*

Citation:
de-Dios-Flores I (2019) Processing Sentences With Multiple Negations: Grammatical Structures That Are Perceived as Unacceptable. Front. Psychol. 10:2346. doi: 10.3389/fpsyg.2019.02346

INTRODUCTION

A central question in the study of sentence comprehension has to do with defining the role that grammatical information plays during the incremental interpretation of language. In this quest, the focus has been placed on studying the sensitivity that language users exhibit to grammatical contrasts during sentence processing. This sensitivity appears to be quite detailed, as instantiated by the skillful accuracy with which language users routinely detect grammatical anomalies both in online experiments and in offline judgments (for reviews, see Kaan, 2007; Phillips et al., 2011; Sprouse et al., 2013; Sprouse and Lau, 2013; Lewis and Phillips, 2015). The grammatical richness with which the language comprehension system seems to operate makes it even more interesting when the outputs of sentence processing do not converge with the constraints of the grammar. Misalignments between grammar and parsing provide a unique window into the principles that guide language comprehension, and their study has been a fruitful program in psycholinguistic research, giving way to numerous models and theories. Such grammar-parser discrepancies have been identified in a variety of structures and are explained by appealing to different grammatical and psychological principles. Without getting into the details of each specific case for reasons of space, the current mosaic of misalignments can be summarized attending to two criteria: first, whether they occur in grammatical or ungrammatical sentences; second, whether they are revealed in *fast* responses (observed in online processing tasks) or they also impact *slow* responses (observed in offline acceptability judgments).

Since its early days, linguistics has subscribed to the relatively uncontroversial view that grammatical sentences may be deemed unacceptable for reasons that are independent of grammatical theory (Chomsky, 1957). Some sentences are – almost – impossible to parse because their complexity exceeds the capacity of the system, leading to processing overload. This is the case of widely studied phenomena like multiple center embedding (e.g., Chomsky and Miller, 1963; Miller and Isard, 1964; Gibson, 1998) or strong garden path sentences (e.g., Bever, 1970; Frazier and Rayner, 1982; MacDonald et al., 1994), illustrated in (1) and (2), respectively.

(1) The patient who the nurse who the clinic had hired admitted met Jack.
(2) The horse raced past the barn fell.

Even though these sentences abide by the constraints of the grammar of English, it has long been known that most native speakers find them incomprehensible, exhibiting great difficulties in processing tasks and judging them as unacceptable in offline ratings. The opposite case can also be found, as certain ungrammatical configurations are sometimes processed and judged as if they were acceptable. So-called comparative illusions, illustrated in (3), are one of the most striking examples of this (Pullum, 2004; Wellwood et al., 2018). When native speakers are presented with sentences like (3), they remarkably judge them as both acceptable and meaningful; and only upon guided examination do they become aware of their ungrammaticality and semantic incoherence. A similar effect is observed when the multiple center-embedded sentences in (1) are presented to speakers with only two verbs instead of the required three, as shown in (4). Whereas the sentence is now ungrammatical, processing measurements and acceptability ratings improve when compared to its grammatical counterpart in (1) (Frazier, 1985; Gibson and Thomas, 1999; Gimenes et al., 2009; Häussler and Bader, 2015). This effect is sometimes referred to as the missing VP illusion. Comparative illusions and missing VP illusions are explained on the basis of different operations but display the same pattern of misalignment that opposes grammatical knowledge with online/offline responses.

(3) *More people have been to Russia than I have.
(4) *The patient who the nurse who the clinic had hired met Jack.

Although sentences like (3) are referred to as a *comparative illusions*, the label *grammatical illusion* is generally used to describe situations in which comprehenders fail to notice a grammatical error in processing tasks but clearly recognize the same sentences as unacceptable in offline judgments (Phillips et al., 2011; Lewis and Phillips, 2015). This is the case of agreement illusions, illustrated in (5) (Bock and Miller, 1991; Pearlmutter et al., 1999; Staub, 2009; Wagers et al., 2009) and negative polarity item illusions, illustrated in (6) and extensively covered in the next section. Despite the ungrammaticality of these examples, online processing measures indicate that the parser initially treats them as correct due to the presence of intervening elements: the plural *cabinets* in (5) and the negative quantifier *no* in (6). That is, grammatical illusions are typically described as discrepancies between fast (online) and slow (offline) responses, implying that online and offline measures of acceptability reflect qualitatively different aspects of linguistic behavior. In the general discussion, we will challenge such a neat view of grammatical illusions, as we hope to show that illusion-like patterns can emerge in the absence of a straightforward contrast between online and offline responses. Furthermore, even though grammatical illusions have attracted much interest in the past few years, the opposite phenomenon (i.e., illusions of ungrammaticality) is less often discussed. This project draws from research on negative polarity item (NPI) illusions in order to explore a candidate structure for illusions of ungrammaticality that illustrated by the grammatical sentence in (7) and explained in detail in section "The Current Investigation: Multiple Negation."

(5) *The key to the cabinets are on the table.
(6) *The bills that no senators voted for have ever become law.
(7) The bills that no senators voted for have never become law.

The heterogeneous inventory of misalignments has motivated a debate about the role that grammatical information plays during sentence comprehension. This debate is embodied in the two systems/one-system divide (Lewis and Phillips, 2015). Proponents of a two system architecture (e.g., Townsend and Bever, 2001; Ferreira et al., 2002; Ferreira and Patson, 2007; Frank et al., 2012; Trotzke et al., 2013) argue that language comprehension and production are supported by a set of heuristic

procedures that do not require speakers to build detailed grammatical information. Under this view, grammar is conceived of as a static body of knowledge that speakers can consult to verify the acceptability of sentences, and misalignments simply reflect the different outputs of these two systems. As Lewis and Phillips (2015) point out, this view is faced with the challenge of explaining how, in the majority of cases, comprehension and production actually exhibit grammatical richness and accuracy. By contrast, the strong convergence between grammar and parsing can be easily explained under a one-system view. In a one-system view, grammar and parsing are understood as forming a single cognitive system that serves the needs of comprehending and producing language (e.g., Phillips and Lewis, 2013; Embick and Poeppel, 2015; Lewis and Phillips, 2015; Mancini, 2018). In this architecture, grammar is an abstract description of the representations that the system builds. Instead of considering misalignments to be arbitrary failures, proponents of the one system view seek to understand the common profile of misalignments in order to systematically predict which linguistic computations will cause the system to err. In this vein, the present work takes NPI illusions as a starting point in order to explore a new and interesting instance of misalignment observed for grammatical sentences like (7). We start by discussing the specifics of NPI illusions that motivate this investigation.

Negative Polarity Item Illusions

NPIs constitute a closed class of lexical items instantiated by words like *ever, any,* or *yet* that tend be used to strengthen the statements in which they appear (Kadmon and Landman, 1993). The heterogeneous nature of the contexts in which NPIs are licensed has motivated a wide range of theories within formal linguistics. These tend to capture the licensing conditions as an interaction of syntactic, semantic and pragmatic mechanisms (Barker, 1970, see Barker, 2018 for a recent review; Ladusaw, 1979; Linebarger, 1987; Krifka, 1995; Giannakidou, 1998, 2011). One of the most prominent licensing environments for NPIs is contexts that have some negative element[1]. For example, in (8a) the NPI *ever* is licensed by the presence of the negative quantifier *no* in subject position, while its absence in (8b/c) renders the sentences ungrammatical. Yet – as becomes apparent in the ungrammaticality of (8b) – mere linear precedence of the negative element is not enough: the NPI must occur in a position in which the negative quantifier is structurally accessible, a condition that is often explained as corresponding to overt c-command (Laka, 1994).

(8) a. **No** authors [that the critics recommended] have **ever** received acknowledgement for a best-selling novel.
b. *The authors [that **no** critics recommended] have **ever** received acknowledgement for a best-selling novel.
c. *The authors [that the critics recommended] have **ever** received acknowledgement for a best-selling novel.
(Parker and Phillips, 2016)

The most interesting property of sentences like (8b) is that comprehenders often fail to notice their ungrammaticality because the presence of the negative quantifier in the relative clause reduces the effects of disruption observed for unlicensed NPIs, such as *ever* in (8c). Even though (8b) and (8c) are equally ungrammatical, processing experiments find (8b) to be parsed with much more ease than (8c). However, illusion effects do not always improve ungrammatical sentences like (8b) on a pair with perfectly grammatical ones. For example, in speeded acceptability tasks, NPI illusions arise as a three-way distinction[2] among the conditions. Importantly, the interference generated by the negative quantifier seems to be only temporary. When participants are given enough time to judge the sentences, both (8b) and (8c) are recognized as unacceptable. This interference effect is known as an NPI illusion, a subtype of illusion of grammaticality. It is empirically robust across languages and measurements, such as speeded acceptability (German: Drenhaus et al., 2005; English: Parker and Phillips, 2016; de-Dios-Flores et al., 2017; Korean: Yun et al., 2018), self-paced reading (English: Xiang et al., 2013; Parker and Phillips, 2016), eye-tracking (German: Vasishth et al., 2008), or event-response potentials (German: Drenhaus et al., 2005; English: Xiang et al., 2009; Turkish: Yanilmaz and Drury, 2018).

Initial accounts of NPI illusions explore two different licensing routes debated in the grammar of NPIs as the source of the effect. On the one hand, Vasishth et al. (2008) propose that the interference effect arises as the consequence of retrieving the irrelevant licensor *no* due to partial feature match in a cue-based memory architecture (Lewis and Vasishth, 2005). This account rests on the assumption that NPI licensing involves establishing a direct item-to-item dependency between the NPI and a grammatical licensor using semantic (i.e., [+negative]) and syntactic (i.e., [+c-command]) cues. Thus, partial match with one of the two cues (i.e., [+negative]) would generate the acceptability illusion. However, it has been argued that NPIs can also be licensed pragmatically through negative inferences (Linebarger, 1987; Giannakidou, 2006). Building on this idea, Xiang et al. (2009, 2013) proposed instead that illusory licensing could be the result of generating negative inferences about the contrasting set of referents denoted by the relative clause in (8b), that is, *the authors that the critics recommended,* which would not have the predicated property (i.e., *receive and acknowledgment*). According to this proposal, these erroneous negative inferences could produce the licensing illusion. While these two accounts appeal to different grammatical resources available to license NPIs (syntactic-semantic vs. pragmatic), they both explain the intrusion effect by the misapplication of the licensing mechanisms activated when encountering the NPI. Accordingly, the two views predict, in broad terms, that illusions should generalize to different items and configurations whenever an NPI has to be licensed.

Nonetheless, a more recent investigation by Parker and Phillips (2016) has provided compelling experimental and modeling

[1]Questions, conditionals, or comparative structures are also frequent licensing environments (see Giannakidou, 2011, for a recent review).

[2]Given that speeded acceptability tasks present a two-alternative forced choice (yes/no), the three-way distinction arises because not all the participants experience an allusion for all the items.

evidence that the configurations that yield NPI illusions are more restricted than it was initially thought. In a series of experiments, they demonstrate that the intrusion effect can be turned off by increasing the distance between the NPI and the illicit licensor as in (9)[3] or (10). This behavior is not predicted by previous accounts. Parker and Phillips (2016) argue that the *on/off* behavior of NPI illusions points to changes in the status of the encoding that is probed for licensing at the point of dependency formation, emphasizing the idea that linguistic encodings are not stable but, rather, take some time to complete. Consequently, NPI illusions reflect access to intermediate stages of the encoding process. When the NPI is checked against the licensing context soon after the relative clause has been encountered, the irrelevant negation may be transparently accessible to spuriously license the NPI. However, when the encoding of the licensing context is accessed at a later point in time, as in (9) and (10), the material inside the relative clause is – presumably – fully encoded and no longer accessible for licensing. This proposal will be referred to as the *changing encodings hypothesis*. Even though it focuses on memory encoding mechanisms rather than retrieval ones, this view is presented as compatible with a cue-based parsing architecture. Putting together ideas from Vasishth et al.'s (2008) proposal with other parsing models that do assume that the format of representations changes over time (e.g., tensor-product variable bindings or vector-based models), Parker and Phillips (2016) speculate that NPI illusions could result from a two-stage representation building process: during a first stage, individual feature values – such as negation – are thought to be transparently accessible giving way to partial match interference. Thus, the licensing illusion could occur during this first stage. In the second stage, individual features are thought to be bound together into a distributed representation, and they could no longer be independently evaluated, blocking illusions to happen. Such an explanation can account for the presence of interference in sentences like (8a) and the absence of it in sentences like (9) and (10).

(9) The authors [that **no** critics recommended] have received **any** acknowledgement for a best-selling novel.
(10) The authors [that **no** editors recommended] have, as the editor mentioned, **ever** received a pay raise.

An alternative speculation about NPI illusions, which will be referred to as the *ever-never confusability hypothesis*, has not been explicitly maintained or experimentally tested before, but it is briefly discussed by Parker and Phillips (2016, pp. 227–228). This proposal hypothesizes that a confusion between *ever* and *never* could be behind the improved perception of NPI illusion sentences. Such a confusion is thought to be facilitated by the orthographic and phonological similarities of the two words, and crucially, because substituting *ever* with *never* would provide a grammatical continuation for NPI sentences like (8b). A process of this sort can be conceptualized under a noisy-channel architecture of sentence comprehension (Levy, 2008a,b; Levy et al., 2009; Gibson et al., 2013). Noisy-channel models assume that retaining each individual word in short-term memory introduces a degree of uncertainty about the previous input. When processing problems are encountered, this uncertainty gives rise to the possibility of misrepresenting previous words in the sentence in cases in which a near neighbor would allow a more probable structure and/or repair an error in the input. For the case of NPI illusions, uncertainty about the input is expected to increase when comprehenders encounter an unlicensed NPI, causing *ever* to be misrepresented in a proportion of cases as its near neighbor *never*, repairing the ungrammaticality[4]. But, why would comprehenders misrepresent the input for sentences with an irrelevant licensor (8b) and not for sentences with no licensor at all (8c)? A possible explanation is that *never* is actually a more plausible continuation for sentences containing the negative quantifier in the relative clause than it is for sentences without it. If we take the examples in (8), it is easier to conceive a situation in which a set of authors have never received acknowledgement when they were not recommended by the critics (8b), than when they were recommended by the critics (8c). Consequently, *ever* could be more often misrepresented as *never* in (8b) than in (8c), explaining the improved perception of NPI illusion sentences. The present investigation explores the *changing encodings* and the *ever-never confusability hypothesis* by examining the processing and acceptability of sentences in which the NPI *ever* was substituted by the negative adverb *never.* Sections "The Current Investigation: Multiple Negation" to "Predictions: Relating Multiple Negation to Negative Polarity Item Illusions" present the details of the experimental design and the specific predictions on which it is articulated.

The Current Investigation: Multiple Negation

The experiments presented here make use of different configurations of negative elements as a means to test two contrasting predictions inspired by previous accounts of NPI illusions. For this purpose, this investigation focuses on grammatical sentences, which vary the presence and structural location of the negative determiner *no* with respect to the adverb *never.* This manipulation results in the three contrasts shown in **Table 1**: single negation (condition A), multiple negation (condition B), and double negation (condition C). The main objective of the project is to study the processing and acceptability of multiple negation sentences (condition B). In these sentences, the negative adverb *never* is linearly preceded by a structurally inaccessible negation, the quantifier *no* inside the relative clause. Multiple negation sentences could be considered the opposite configuration of NPI illusions in that when the NPI *ever* is substituted by *never,* they become

[3]Note that the post-verbal placement of *any* is enough distance in order to turn the illusion off.

[4]As noted by an anonymous reviewer, under a noisy-channel architecture, it is possible that *never* is misrepresented as *ever* in cases in which the later allows a more probable structure (e.g., *No one never came*). Nonetheless, as an explanation of NPI illusions, the *ever-never confusability hypothesis* is not assumed to go both the ways. If substituting *ever* with *never* is what is thought to improve the perception of the unlicensed NPI *ever* in illusion sentences, then *never* is assumed to be parsed without problems.

TABLE 1 | Sample set of experimental conditions.

A. Single negation	**The** authors [that **the** critics recommended] have **never** received acknowledgment for a best-selling novel.
B. Multiple negation	**The** authors [that **no** critics recommended] have **never** received acknowledgment for a best-selling novel.
C. Double negation	**No** authors [that **the** critics recommended] have **never** received acknowledgment for a best-selling novel.

grammatical strings. More importantly, as noted in Huddleston and Pullum (2002), the words *ever* and *never* are semantically and etymologically related: both elements express a quantification in terms of frequency or temporal location, despite having different syntactic distributions. While the NPI *ever* adds a quantificational force to an already negated statement, *never* expresses the negative and quantificational forces at the same time. Thus, sentences like *No authors have ever received acknowledgement* and *The authors have never received acknowledgement* are roughly equivalent.

In order to study the processing and acceptability of multiple negation sentences, they will be compared to single negation and double negation sentences using three tasks: a speeded acceptability task (Experiment 1), a self-paced reading task (Experiment 2), and an untimed acceptability judgment (Experiment 3). The first two experiments are devised to tap into the online/fast processing of these structures, while Experiment 3 focuses on speakers' offline/slow perception of acceptability. Importantly, the three experimental conditions tested here are grammatical in English, even though they vary in their syntactic and semantic complexity. For the purpose of this investigation, single negation sentences are taken to be the simplest of the three and serve as an unproblematic baseline for comparison. On the other end, instances of double negation are assumed to generate processing and acceptability problems, and are used as some sort of "unacceptable" or degraded baseline. These initial assumptions are based on previous linguistic considerations, which are reviewed in the next section.

Some Notes on Negation

All natural languages express negation (Horn, 2001). Yet, in spite of the high frequency with which negative expressions appear in routine language use, negative statements have been related to an increase in processing effort when compared to equivalent affirmative statements (Wason, 1961; Fischler et al., 1983; Carpenter et al., 1999; Kaup et al., 2006; Herbert and Kübler, 2011). There is vast cross-linguistic variation on how the operation of negation can be carried out with regard to "the position of negative elements, the form of negative elements and the interpretation of sentences that consist of multiple negative elements" (Zeijlstra, 2007, p. 498). In English, negation can be marked by words (e.g., *no, not* or *never*) or by affixes (e.g., *-n't* or *in-*). For instance, in the single negation condition in **Table 1**, the negative adverb *never* expresses sentential negation. As regards the presence of more than one negative element in a sentence, one can often find sentences composed of two clauses that are independently negated. This case is illustrated by multiple negation sentences. Zeijlstra (2004, p. 58) points out that these sentences should not be considered as double negation because "two propositions are negated one, but no proposition is negated twice". To avoid confusion, the label *multiple negation* will only be used to refer to these sentences.

Furthermore, when two negative expressions interact in the same clause, they can form two types of dependencies: negative concord or double negation. Negative concord dependencies are observed in languages in which the presence of two negative elements is interpreted as a single semantic negation (e.g., Spanish, Italian, or African-American Language). Conversely, Standard English is commonly classified as a double negation language, in which each negative marker contributes an independent semantic negation. In double negation languages, the two negative elements cancel each other out yielding an affirmative interpretation as a result (Horn, 2001, 2010; de Swart, 2010; Puskás, 2012 i.a.). This is exemplified by the double negation sentence in **Table 1**, which could be paraphrased as *All the authors that the critics recommended have received acknowledgement for a best-selling novel at least once*. Double negation dependencies entail complex operations in terms of the syntactic, semantic and prosodic marks that are needed. For instance, it has been found that the use of specific contradictory intonational contour and denial gesture features are crucial for the felicitous interpretation of double negation dependencies in oral comprehension tasks (Espinal and Prieto, 2011; Prieto et al., 2013). In written format, a corpus study by Larrivée (2016) described that double negation dependencies are generally triggered in restricted information-structure configurations in which a discourse-old negative statement is being denied by the second negation. Due to its complexity, double negation dependencies are assumed to engage in greater processing cost than negative concord dependencies or single negation (Corblin, 1996). Unfortunately, the psycholinguistic studies on the processing of double negatives are very scarce.

Using a sentence verification task, Sherman (1976) tested multiple combinations of negative elements (from 1 up to 5 negative markers).[5] His results clearly show that the presence of two negative elements in a sentence considerably increases comprehension time and error rates. Another study by Schiller et al. (2017) used the Event-Related Potential technique in order to study configurations that combine verbal negation and affixal negation (e.g., *not impossible*). Their findings show that, at least for these simpler combinations, the processing disruptions associated with double negation can be overruled by discourse contexts that clearly evoke negative expectations. Putting the evidence from these previous studies together, it appears that instances of double negation seem to present parsing difficulties when there are not explicit pragmatic cues that help predict the double negative dependency. Along these lines, Blanchette (2013, see also Blanchette et al., 2018) maintains that speakers of Standard English tend to interpret instances of double negation as negative concord dependencies when they are encountered in the absence of the relevant cues. This claim is also supported by experimental evidence

[5]The materials included combinations of sentential negation with affixal and semantic negation (e.g., *not, no one, doubted, un-*).

provided by Thornton et al. (2016), showing that young children acquiring Standard English initially perceive double negation configurations as forming negative concord dependencies. Taking all this into account, our starting assumption is that the double negation dependencies used here (i.e., condition C, **Table 1**) will generate strong processing difficulties and will be deemed unlicensed when encountered in isolation. By contrast, single negation (i.e., condition A, **Table 1**) is expected to be processed with ease and to be recognized as acceptable. These assumptions are set to test in the experiments that follow, and their endorsement is essential in order to interpret them as baselines. Before moving into the experimental evidence, the next section discusses the specific predictions that relate multiple negation sentences to NPI illusions.

Predictions: Relating Multiple Negation to Negative Polarity Item Illusions

Given that the grammar of *never* is not constrained by the licensing conditions that affect NPIs, this investigation does not address explanations of NPI illusions that invoke the faulty implementation of NPI-specific licensing mechanisms (Vasishth et al., 2008; Xiang et al., 2009, 2013). The interest of this project lies, instead, on exploring two conflicting predictions that can be extracted from the *ever-never confusability* and the *changing encodings* hypotheses. Multiple negation sentences provide a ground for testing these two proposals because they predict opposite patterns of results.

On the one hand, if NPI illusions are the result of misrepresenting *ever* as *never*, multiple negation sentences display the precise configuration that is assumed to rescue unlicensed NPIs. In a noisy-channel architecture, comprehenders would be likely to misinterpret *ever* as *never* in cases in which *never* provides a more plausible and/or natural sentence. Here, a correspondence is assumed between plausibility and grammaticality, as sentences containing *never* are thought to be more plausible because they provide a grammatical and meaningful continuation. This hypothesis predicts that multiple negation sentences should be recognized as acceptable by native speakers of English and should be parsed with ease. If multiple negation sentences are perceived as acceptable, they are expected to pattern closer to single negation sentences (which are assumed to be perceived as acceptable) than to double negation sentences (which are assumed to generate problems). Importantly, these predictions result from our understanding of the *ever-never confusability hypothesis* within a noisy-channel architecture, as this proposal had never been explicitly articulated until now. In our view, an explanation that appeals to a misrepresentation of *ever* as *never* is in conflict with multiple negation sentences generating strong processing or acceptability problems; because such a misrepresentation is only motivated when it leads the parser into an acceptable and unproblematic structure. The *ever-never confusability hypothesis* – or any other account of NPI illusions – does not predict sentences like (8b) to be perceived on a pair with perfectly grammatical ones. Yet, this hypothesis rests on the assumption that similar sentences containing *never* should be generally processed and recognized as acceptable.

On the other hand, the *changing encodings hypothesis* put forth by Parker and Phillips (2016) predicts the opposite outcome. Under this view, NPI illusions are the result of accessing incomplete computations of the material inside the relative clause that includes the quantifier *no*, facilitating a dependency between the spurious licensor and *ever*. The negative quantifier could be retrieved as a licensor – possibly in a cue-based procedure – because its individual features can be transparently accessible in early stages. Accordingly, the same intermediate stage computations are expected to be in place in multiple negation sentences up to the point when participants reach *never*. If the negation inside the relative clause has not been bounded into a distributed representation when *never* is encountered, the parser may experience problems in having to integrate two negative elements simultaneously. This cost is predicted to manifest as a disruption in reading times in multiple negation sentences relative to single negation. Importantly, the content that precedes *never* in multiple negation sentences is the same that precedes *ever* in Parker and Phillips' (2016) illusion configurations. Thus, finding similar interference effects in online tasks could indicate that the same incomplete encodings that temporarily improve the perception of ungrammatical NPI configurations are responsible for hampering the comprehension of grammatical multiple negation sentences. A possible speculation is that the disruption predicted in multiple negation sentences could index the parser's difficulties evaluating a double negation dependency between *no* and *never*. Assuming that double negative dependencies are problematic in the absence of enough contextual cues, entertaining an illusory double negation dependency is expected to generate similar effects to those expected in actual double negation sentences.

In sum, the fundamental question that this research aims to answer is whether multiple negation sentences are processed and judged closer to single negation sentences (which are expected to be processed without any problems), or to double negation sentences (which are expected to generate strong disruptions). Moreover, although this investigation takes NPI illusions as a point of departure, we hope that it will also provide insights into the processing and grammatical status of double negation dependencies; a phenomenon that still remains poorly explored from a psycholinguistic perspective.

EXPERIMENT 1: SPEEDED ACCEPTABILITY JUDGMENT

Experiment 1 used the speeded acceptability technique to investigate whether the perception of grammatical sentences containing two negation markers is degraded for sentences in which these negative elements do not engage in a negative dependency. Speeded acceptability judgments are generally considered an *online* technique because the limited amount of time provided to respond forces participants to operate on fast and unconscious intuitions of grammaticality. They have been reliably used as a time-sensitive measure to test NPI illusion configurations (e.g., Drenhaus et al., 2005; Parker and Phillips, 2016; de-Dios-Flores et al., 2017; Muller et al., 2019).

Participants

Twenty-eight native speakers of English (19 female, mean age 20 y/o) participated in this experiment. They were recruited through the University of Maryland's participant database. Participation was compensated with a credit in an introductory linguistics course or, alternatively, with $10. The speeded acceptability task was administered together with another unrelated experiment as part of a 1-h testing session. All the participants in this, and the following experiments provided informed consent and were naïve to the purpose of the experiment. They were also screened for native speaker abilities through a short questionnaire that tested constraints on tense, modality, morphology, ellipsis, and syntactic islands. In order to participate in the experiments, the candidates were required to pass the test with a minimum of 7/9.

Materials

The experimental materials consisted of 36 sets of three items like those in **Table 1**. These were adapted from the stimuli used in Parker and Phillips (2016) by solely substituting the NPI *ever* by the negative adverb *never*. The experimental conditions were counterbalanced in three lists using a Latin Square design, together with 72-filler sentences of similar internal structure, length and complexity. Each list had a total of 108 items, and participants were randomly assigned to one of the three lists. Grammaticality was balanced so that half of the sentences were ungrammatical. This ensured that the initial probabilities of providing a *yes* or a *no* answer were equal across the task. For this purpose, double negation sentences (condition C) were counted as ungrammatical. To achieve a 1:1 grammatical-to-ungrammatical ratio, 42 fillers contained ungrammaticalities. The grammatical violations introduced included preposition usage, number agreement, verbal morphology and pronoun-antecedent mismatches. During the delivery of the instructions, participants were asked to complete six practice items to ensure that they had understood the procedure.

Procedure

The stimuli for this speeded acceptability task were presented on a desktop PC using Ibex (Drummond, 2013). Each sentence was displayed word by word at a rate of 400 ms per word, in the center of the screen, using the rapid serial visual presentation (RSVP) paradigm. At the end of each sentence, a response screen appeared and participants were asked to provide a *yes*/*no* button press judgment in a maximum time of 2 s. When participants failed to provide the judgment in time, a message indicated that they were too slow. Participants were instructed to read the sentences carefully and judge whether they came across as well-formed English. They only received feedback for the first two practice items. All participants were tested on the same computer. The task lasted for approximately 30 min, and the order of presentation for experimental and filler sentences was randomized for each participant.

Analysis

The *yes/no* responses collected were analyzed using a generalized linear mixed model for binomial distributions (also known as logistic mixed model; Baayen et al., 2008; Jaeger, 2008). A maximal model with a fully specified random effects structure was initially built. This model included the experimental conditions as fixed effects and by-participants and by-items random intercepts and slopes for the experimental conditions. Yet, this model failed to converge and had to be reduced to a model with random intercepts but no slopes. This was the maximally converging model. As noted in Barr et al. (2013, pp. 23–24), categorical data tend to pose more difficulties for maximal models to converge. For this and the following two experiments, the contrasts among the three experimental conditions were obtained as follows: first, condition A (single negation) was used as the reference level of the intercept in order to obtain the contrasts between A and B (multiple negation) and A and C (double negation). Then, the contrasts between B and C were obtained by setting B as the intercept. All the analyses reported for this and the following experiments were carried out using R, an open-source programming environment for statistical computing (R Development Core Team, 2014). Specifically, the models were estimated using the package *lme4* for linear mixed effects models (Bates et al., 2015). Following Gelman and Hill (2006), an effect was considered statistically significant at the $p > 0.05$ level when the absolute z value was above 2.

Results

Figure 1 shows the average percentage of *yes* responses to each of the three experimental conditions. Sentences containing a single negation (condition A) were judged as acceptable in most cases (above 80% acceptance). The presence of two negations significantly reduced the perception of acceptability for both multiple negation (A vs. B: $\hat{\beta} = -1.48$, SE = 0.21,

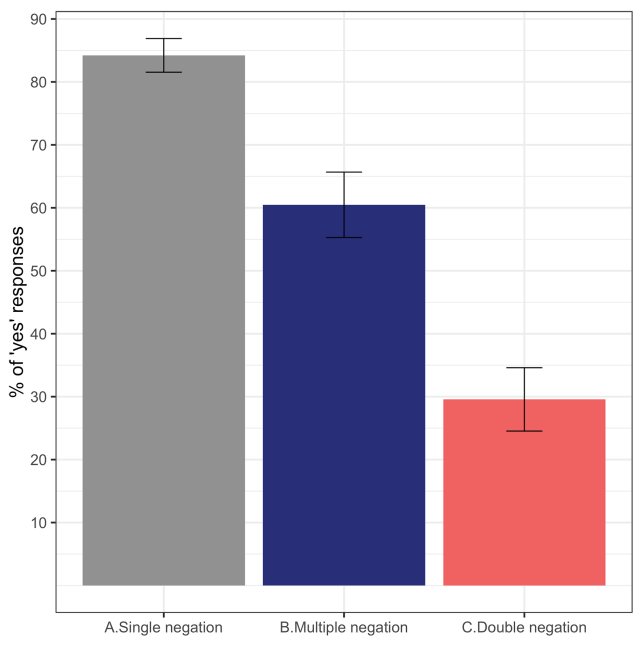


FIGURE 1 | Average percentage of "YES" responses for the experimental conditions aggregated by participant (Experiment 1). Error bars indicate standard error of the mean.

$z = -7.06$) and double negation sentences (A vs. C: $\hat{\beta} = -3.05$, SE = 0.23, $z = -13.35$). Nonetheless, the decrease in acceptability was less acute when the two negations appeared in different clauses (condition B, above 60% acceptance) than for traditional double negatives (condition C, below 30% acceptance). This contrast was statistically significant (B vs. C: $\hat{\beta} = -1.57$, SE = 0.19, $z = -8.31$).

Discussion

This experiment tested the impact of different negation configurations on fast perceptions of acceptability using a processing demand task. Based on linguistic and psycholinguistic considerations, it was initially assumed that single negation sentences would be unproblematic for native speakers of English, and that double negation sentences would possibly be deemed unacceptable in the absence of the appropriate licensing context. These assumptions are borne out in the results. Importantly, the perception of acceptability of grammatical multiple negation sentences is penalized, although there is a significant three-way distinction among the conditions: single negation sentences are accepted in the vast majority of cases, multiple negation sentences exhibit a lower but still greater proportion of *yes* over *no* responses, and double negation sentences are rejected in the majority of cases. That is, the perceived ungrammaticality increased when *no* c-commanded *never* than when it did not. The acceptability contrast between single negation and multiple negation sentences cannot be attributed to constraints of the grammar or other linguistic considerations, as both sentences are perfectly grammatical. Instead, it points to a processing problem as the source of the effect. These results are interpreted as initial evidence that the presence of a structurally inaccessible negative quantifier *no* interferes with the integration of the adverb *never* in the main clause. The pattern of results bears significant resemblance to the picture that arises in speeded acceptability studies of NPI illusions (e.g., Drenhaus et al., 2005; Parker and Phillips, 2016; de-Dios-Flores et al., 2017; Muller et al., 2019). In these studies, an illusion of grammaticality is identified with higher acceptance rates for unlicensed NPIs in sentences with *no* inside the relative clause (e.g., *The authors [that no critics recommended] have ever received acknowledgement for a best-selling novel*), compared to sentences without it. The pattern found in this experiment is the exact opposite: lower acceptance rates for grammatical sentences with *no* inside the relative clause (i.e., multiple negation) than for similar grammatical sentences without it (i.e., single negation). In other words, while the intrusive *no* ameliorates the perception of *ever* in ungrammatical sentences, it seems to deteriorate the perception of *never* in grammatical sentences.

Speeded acceptability tasks gather information about the participants' overall initial perception of acceptability, and they have been proved to be a reliable technique uncovering grammatical illusions. Even though participants respond once the full sentence has been presented, the proportion of correct judgments is generally assumed to relate to processing operations due to the time pressure under which these are elicited. However, as Vasishth et al. (2008, pp. 696–697) point out, "the source of the judgment itself is presumably a decision process that takes as input the products of (possibly partially) completed online processing." Thus, this task does not allow us to ascertain which are the specific sentence regions that generate this behavior or to disentangle sentence comprehension mechanisms from other processes that affect end-of-sentence decisions. The next experiment was designed to delve deeper into the source of the interference effect.

EXPERIMENT 2: SELF-PACED READING

This experiment uses a self-paced reading task in order to study the online processing of the sentences under investigation. This method provides access to moment-by-moment processing during the automatic integration of each sentence word and the difficulty generated by it. In light of Experiment 1, the integration of *never* is expected to take place without problems in single negation sentences and to generate strong processing disruptions in double negation sentences. With regard to the critical condition (i.e., multiple negation sentences), if the presence of the negative quantifier *no* in the relative clause interferes with the integration of *never* in the main clause, longer reading times are expected at the point of *never* in multiple negation sentences relative to single negation sentences.

Participants

The participants in this experiment were 36 native speakers of English (30 female, mean age 24 y/o) who were recruited in the area of Santiago de Compostela. All of them were pursuing or had just finished university education (BA or MA) in different disciplines in the USA and were serving as high school language assistants as part of a 1-year exchange program funded by the Galician Regional Ministry of Education[6]. Special care was taken to ensure that none of the participants had spent more than 48 months outside an English-speaking country across their entire life. Their participation in the study was voluntary.

Materials

The experimental materials used in this task were the same as in Experiment 1 (see **Table 1**). The three conditions were counterbalanced in three lists together with a grammatical version of the same 72-filler sentences. The ratio of ungrammatical-to-grammatical sentences was reduced in order to prevent participants from developing unnatural reading strategies. In order to ensure that participants were reading for comprehension, all the experimental and filler sentences were followed by a *yes*/*no* question. These comprehension questions addressed pieces of information located in different parts of the sentences. This way, participants were forced to pay equal attention to all the sentence regions. The comprehension questions for the experimental items were never related to the negated material and the probability of providing a positive or a negative answer was balanced. During the delivery of the instructions, participants were asked to complete four practice items to ensure that they had understood the procedure.

[6]https://www.edu.xunta.gal/portal/es/linguasestranxeiras/1640/1643

Procedure

The task was implemented in Inquisit 4 (Millisecond Software, 2015) using the non-cumulative word-by-word moving window version of the self-paced reading procedure (Just et al., 1982). In this version of the task, participants are presented with the entire sentence on the screen with each word masked by dashes and separated by spaces. When the predefined key is pressed (the space bar in this case), the first word is revealed. When the space bar is pressed one more time, the second word appears and the first word is re-masked. By collecting the time elapsed between bar-presses, this task allows us to measure the time spent in each word. Participants were instructed to keep their fingers on the selected keys (i.e., the space bar and *yes/no* keys) for the entire session. This way, they could move forward easily at their own pace and answer the questions as accurately and as fast as possible. They received on screen feedback for both wrong and right answers – the word "right" was displayed for 1,000 ms when the response was correct, and the word "wrong" was displayed for 2,000 ms when the response was incorrect. All participants were tested using the same computer. The task lasted for approximately 35 min, and the order of presentation of experimental and filler sentences was randomized for each participant.

Analysis

The acceptance threshold for accuracy in the questions was set to 80% to ensure that the final sample only contained participants that were reading for comprehension. No participant had to be excluded from the analysis due to poor performance. Unrealistic reading times were first deleted following standard practices in the self-paced reading literature (see for example Hofmeister, 2011; Nicenboim et al., 2016). These included RTs above 2,500 ms and below 100 ms, which are possibly the product of spurious delays or erroneous button presses that might obscure the initial stages of model fitting (Baayen, 2008). This procedure resulted in the exclusion of 0.85% of the data across all sentence regions and 0.23% of the data in the regions of interest. Subsequently, the remaining reading times were log-transformed in order to reduce non-normality. Average RTs for the experimental conditions were then compared in four regions of interest: the auxiliary verb before *never*, which signals the end of the relative clause; the negative adverb *never,* which is the critical point at which the different negation configurations are established; and the two next spillover words. These reading times were analyzed using a linear mixed effects model.

Following the same model building procedure as in Experiment 1, the RTs in the four regions of interest were analyzed using the maximally converging model (Barr et al., 2013). The maximal model included the experimental conditions as fixed effects and by-participant and by-item random intercepts and slopes. This model was applied in the pre-critical region, the critical region (*never*), and the first spillover region. In the second spillover region, the maximal model had to be reduced due to convergence problems. This reduced model included by-participant and by-item random intercepts but no slopes. The accuracy for the comprehension questions in the experimental trials was also analyzed. This was done by means of a generalized linear mixed model for binomial distributions (Jaeger, 2008). The maximally converging model included fixed effects for the experimental conditions and only random intercepts for participants and items. An effect was considered to be statistically significant at the level of $p < 0.05$ when the absolute t *or* z value was above 2 (Gelman and Hill, 2006; Baayen et al., 2008).

Results

Figure 2 shows the average word-by-word reading times in log-milliseconds for the three experimental conditions in all the sentence regions. The four regions of interest are highlighted inside a square. The model results for the pre-critical region (the auxiliary *have*) show a significant effect of multiple negation when compared with double negation (B vs. C: $\hat{\beta} = -0.07$,

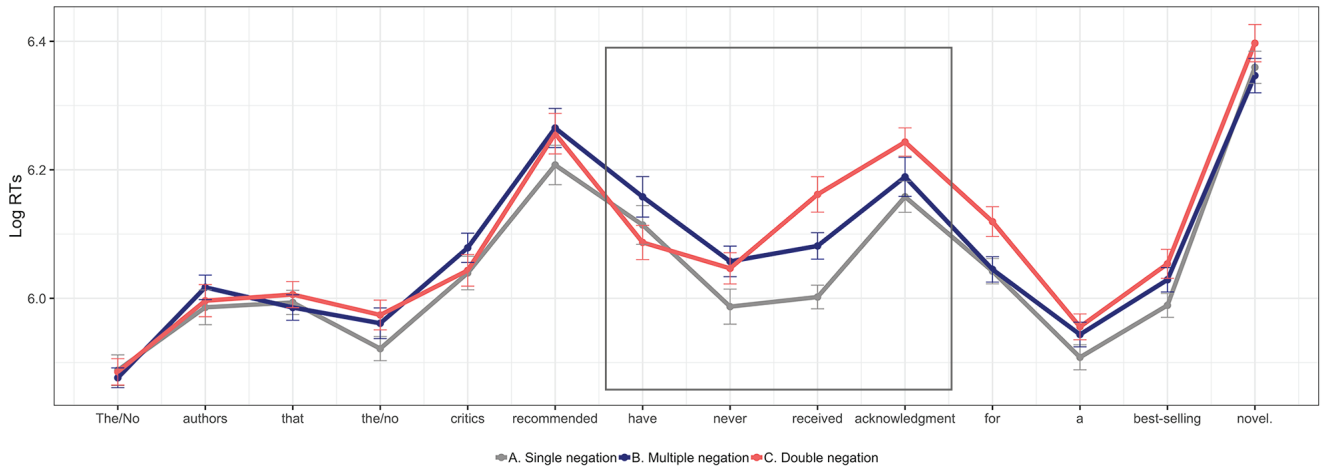


FIGURE 2 | Average word-by-word reading times for the experimental conditions aggregated by participant (Experiment 2). Error bars indicate standard error of the mean. The regions of interest are contained within the square.

SE = 0.03, $t = -2.20$). This region was read slower for multiple negation sentences (condition B) than for double negation sentences (condition C). The contrast with single negation (condition A) was not statistically significant (A vs. B: $\hat{\beta} = 0.04$, SE = 0.03, $t = 1.30$), even though it follows a similar trend. The results for the adverb *never* (the critical region) show that sentences with two negation markers (multiple and double negation sentences) are read more slowly than sentences in which *never* was the only negative element –single negation sentences – (A vs. B: $\hat{\beta} = 0.07$, SE = 0.03, $t = 2.21$; A vs. C: $\hat{\beta} = 0.06$, SE = 0.03, $t = 2.19$). Furthermore, no differences are observed between multiple negation and double negation sentences in the *never* region (B vs. C: $\hat{\beta} = -0.01$, SE = 0.03, $t = -0.37$). The slow-down for sentences with two negations extends to the first spillover region in a three-way contrast: single negation sentences were the fastest of the three (A vs. B: $\hat{\beta} = -0.08$, SE = 0.03, $t = 3.02$; A vs. C: $\hat{\beta} = 0.16$, SE = 0.03, $t = 4.95$). Furthermore, double negation sentences displayed a more pronounced slowdown than multiple negation sentences (B vs. C: $\hat{\beta} = -0.08$, SE = 0.03, $t = 2.88$). In the second spillover region, there was a significant effect of double negation, reflecting slower reading times relative to both single negation and multiple negation sentences (A vs. C: $\hat{\beta} = 0.08$, SE = 0.03, $t = 3.24$; B vs. C: $\hat{\beta} = 0.05$, SE = 0.03, $t = 2.12$). No differences are observed between single negation sentences and multiple negation sentences in this second spillover region (A vs. B: $\hat{\beta} = 0.03$, SE = 0.03, $t = 1.11$). Average accuracy for the comprehension questions in the experimental items was 94% (condition A: 96%, condition B: 96%, condition C: 91%). The results from the logistic regression indicate a significant decrease in accuracy for double negation sentences when compared to the other two conditions (A vs. B: $\hat{\beta} = -0.02$, SE = 0.38, $z = 0.95$; A vs. C: $\hat{\beta} = -0.95$, SE = 0.32, $z = -2.93$; B vs. C: $\hat{\beta} = -0.92$, SE = 0.32, $z = -2.89$).

Discussion

This experiment used the self-paced reading technique to investigate whether the interference effects found in Experiment 1 reflect difficulties in the integration of *never* during the incremental processing of multiple negation sentences. Before examining the results for this critical condition, it is important to note that the reading times for the baseline conditions are aligned with the initial predictions as well as with the results from Experiment 1: single negation sentences are read the fastest of the three, and double negation dependencies did not only impact reading times but also caused a reduction in comprehension question accuracy. This drop in accuracy could be initially surprising because the comprehension questions never targeted information related to the negations and were the same in the three experimental conditions. A plausible explanation for this behavior is that the confusion generated when participants tried to interpret double negated sentences prevented them from paying enough attention to the rest of the contents. The reading times for multiple negation sentences in the regions of interest seem to confirm the intuition that the decrease in acceptance observed in Experiment 1 could arise from the difficulty of integrating *never* when it is preceded by the embedded negative quantifier *no*. Importantly, when comprehenders reach the negative adverb, the reading times for multiple negation and double negation sentences are on a par. The disruption observed for sentences with two negations spills over the sentence regions following *never*, even though participants recover earlier in multiple negation than in double negation conditions. The pattern of results found in this experiment seems initially incompatible with the hypothesis that NPI illusions arise due to a misrepresentation of *ever* as *never.* The *ever/never confusability hypothesis* rests on the assumption that sentences with *never* are both an acceptable and natural continuation, but multiple negation sentences are shown to create processing problems. Such problems are not expected if multiple negation sentences represent the configuration that is thought to ameliorate NPI illusions. The fact that the RTs at the critical region show the same slowdown in both multiple and double negation sentences is particularly relevant because this is the region in which unlicensed NPIs such as (8b) display the strongest facilitation effects. Nonetheless, it is difficult to map the RTs in this experiment to those in classic NPI illusion sentences because of the different baselines used. The offline ratings from the next experiment will hopefully clarify the perceived status of multiple negation sentences.

One potential concern with these results is that the reading times for multiple negation sentences are slower than the other two conditions in the region preceding *never.* Up to this point, the sentences used here are identical to those in Parker and Phillips' (2016) self-paced reading task (Experiment 3 in their work), but they do not observe any significant effects in the pre-critical region. In spite of the lack of statistical contrasts, Parker and Phillips' (2016) data display a similar trend: the auxiliary *have* is read slower in sentences containing *no* inside the relative clause. Given that our sample contained 50% more participants than Parker and Phillips' experiment ($n = 24$), we believe that the two pre-critical effects could be qualitatively similar, but their study lacked the necessary power to detect the contrast. It is also possible that the effect found at the pre-critical region is stronger in our data as a consequence of the experimental manipulation. The auxiliary *have* provides a structural cue that signals the end of the relative clause, and it is always followed by the critical region – *ever* in NPI illusions and *never* in these sentences. However, in the study by Parker and Phillips, the presence of the negative quantifier facilitated the integration of *ever,* while here, its presence seems to hamper the integration of *never.* As the experiment unfolds, the problems associated with the different configurations of negative elements could have made both the quantifier *no* and the adverb *never* more salient in our experiment, and thus, participants could be placing more resources to process the negative quantifier inside the relative clause before reaching the negative adverb. Such an effect is predicted to surface as a slowdown only in multiple negation sentences, as it is the only condition that displays an embedded negation. As suggested by an anonymous reviewer, this conjecture predicts the effect to grow across the experiment, and thus, it can be investigated by modeling the interaction with trial order. However, the results from a post-hoc analysis clearly showed the opposite:

the contrast between multiple negation sentences and the other two conditions was the strongest during the first trials and shrank dramatically across the task[7], discarding this additional possibility. In sum, the fact that the pre-critical effects only arise in multiple negation sentences is interpreted as evidence that at least some aspects of the embedded negation are still being integrated when participants reach the auxiliary *have*. That is, the difficulty associated with the integration of the negative quantifier seems to spill over outside of the relative clause.

If the quantifier *no* has not been fully encoded when comprehenders reach *never,* the slower reading times observed for multiple negation sentences could reflect the difficulties of the parser when trying to integrate two active negative elements. In order to support this interpretation, it is essential to establish whether the contrast observed at the critical and post-critical regions between single negation and multiple negation sentences is not simply the consequence of the spillover effect from the embedded negation. In other words, that there is some additional processing difficulty specifically triggered by *never.* To explore this issue, we calculated Cohen's delta (d) statistic (Cohen, 1988) for the contrast between single and multiple negation sentences in the pre-critical, the critical and the post-critical regions. The results show that whereas the effect size in the pre-critical region is quite small (d = 0.12), the effect size in the critical region is almost three times bigger (d = 0.34), and even more so in the post-critical region (d = 0.51). The fact that the effect grows when *never* is encountered represents evidence that the negative adverb contributes its own source of processing difficulty, and thus, that the disruption observed at the critical and post-critical regions could be reflecting the combined difficulty of integrating the two negative elements. Such interpretation of the results aligns with Parker and Phillips' (2016) hypothesis that NPI illusions arise as a consequence of unstable encodings available when the NPI is being licensed. Under this hypothesis, the slow reading times observed at the negative adverb would reflect the difficulties of the parser to integrate *never* in the context of a previous negative element. As it was speculated in the predictions section, such a disruption could be indexing initial attempts of the parser to entertain a temporary double negation dependency between *never* and *no.* This idea is motivated by the fact that the RTs at the critical region are equally slow for multiple and double negation sentences. The crucial difference between these two conditions is that, in multiple negation sentences, this dependency is not structurally supported, and this could be interpreted as an illusion of ungrammaticality. Two facts seem to support the idea that such an illusory double negation dependency could just be temporarily entertained. First, participants recover earlier from the disruption produced in multiple negation sentences than in double negation sentences. Second, this interference does not seem to have interpretive consequences, inasmuch as comprehension question accuracy is not reduced in multiple negation sentences. The general discussion delves deeper into this issue.

Together, Experiments 1 and 2 provide clear evidence that the negative quantifier *no* inside the relative clause interferes with the online integration of *never* in the main clause. In line with Parker and Phillips' (2016) account of NPI illusions, this interference effect is expected to arise during early parsing stages in which the encodings of the material in the relative clause, and the quantifier *no* in particular, have not been fully computed. Under the assumption that comprehenders only experience an illusion of ungrammaticality in online tasks, native speakers of English are expected to recognize multiple negation sentences as acceptable when given ample time. The objective of Experiment 3 is to test the offline perception of acceptability of sentences under investigation.

EXPERIMENT 3: OFFLINE ACCEPTABILITY RATING

This section presents the results from an offline acceptability judgment (Cowart, 1997). As explained above, acceptability measures will contribute to understand the causes and interpretation of the disruption observed for multiple negation sentences in Experiments 1 and 2. In addition, these untimed ratings will further corroborate the grammatical status of the baseline conditions.

Participants

Twenty-four US-based native speakers of English (6 female, mean age 35 y/o) participated in this experiment. All participants provided informed consent and they received $3 as compensation. The experiment lasted approximately 20 min. Participants were recruited using Amazon's Mechanical Turk (AMT; https://aws.amazon.com/mturk). AMT is a crowdsourcing web-service through which institutions and companies can recruit participants for human intelligence tasks. Its use in the fields of linguistics and psychology has increased in recent years, and several studies have already validated its use for many classical psychological experiments, including tasks using timing measurements (e.g., Crump et al., 2013; Enochson and Culbertson, 2015). For the specific case of acceptability ratings, a large-scale comparison between laboratory-based and AMT-based acceptability ratings conducted by Sprouse (2011) concluded that acceptability data collected in AMT are almost indistinguishable from laboratory data (see also Gibson et al., 2011).

Materials

The materials used in this task were the same 36 sets of experimental items and 72-filler sentences that were used in Experiment 1. The ratio of grammatical-to-ungrammatical sentences was balanced so that half of the sentences across the task contained ungrammaticalities. During the delivery of the instructions, participants were asked to complete six practice items to ensure that they had understood the procedure.

[7]It is relevant for the discussion to report that when trial order is included in the model, the contrast between conditions A (single negation) and B (multiple negation) in the pre-critical region emerged statistically significant. This contrast had not reached statistical significance in the model results reported for Experiment 2.

Procedure

The stimuli were delivered using Ibex (Drummond, 2013). Participants were presented with the entire sentence in the middle of the screen along with a rating scale. Each sentence was presented in an individual screen and participants could only move to the next one once they had emitted a rating by clicking on the scale numbers or, alternatively, using the numbers on their keyboard. Participants were instructed to rate the sentences according to their acceptability in a 7-point scale in terms of whether they came across as well-formed English: 7 meaning totally acceptable and 1 totally unacceptable. In order to help them adjust to the scale, the first two practice items were followed by feedback on "the rating that most people would give in that case" (1 or 2 for an ungrammatical example and 6 or 7 for a grammatical one). They were encouraged to take as much time as they needed and to use the entire range of the scale. The order of presentation of experimental items and fillers was randomized for each participant. The task was completed by all participants in less than 30 min.

Analysis

The ratings collected were analyzed using a linear mixed-effects model that included the experimental conditions as fixed effects and participants and items crossed as random effects. A maximal model with a fully specified random effects structure was initially built. This model failed to converge and the random structure was simplified following Barr et al. (2013). The results reported in the next section correspond to the model with the maximal converging random effects structure, which included by-subject and by-item random intercepts and slopes but no correlation parameters for the by-item grouping. Using a log-likelihood ratio test, this model was compared to a simpler model containing only random intercepts. The test revealed that the maximally converging model provided a better fit to the data ($X^2_{(11)} = 72.37$, $p < 0.0001$). An effect was considered to be statistically significant at the level of $p < 0.05$ when the absolute t value was above 2 (Gelman and Hill, 2006; Baayen et al., 2008).

Results

The results from this experiment are presented in **Figure 3**. Single negation sentences had the highest average rating and double negation sentences the lowest. The acceptability of multiple negation sentences was rated quite low (means: A = 5.66, B = 3.63, C = 2.89). The model results revealed statistically significant differences among the three experimental conditions (A vs. B: $\hat{\beta} = -2.02$, SE = 0.18, $t = -11.15$; A vs. C: $\hat{\beta} = -2.76$, SE = 0.28, $t = -9.93$; B vs. C: $\hat{\beta} = -0.74$, SE = 0.21, $t = -3.46$).

Discussion

The first thing to note about these results is that they confirm the grammatical status that was initially assumed for the baseline conditions: while single negation sentences were judged as perfectly grammatical, double negation sentences were highly rejected. This result is unsurprising in light of Experiments 1 and 2, and it also coincides with the low acceptability ratings reported for double negation sentences in Blanchette (2015). Nonetheless, Experiment 3 was mainly designed to test whether native speakers of English recognize multiple negation sentences as acceptable in spite of the attested processing problems they generate. If these grammatical sentences were recognized as such, the ratings attributed to them should approach those of single negation sentences. However, the results from this task confirm the opposite: the perception of multiple negation sentences is highly degraded compared to single negation sentences. Multiple negation sentences patterned closer to double negation ones, although mean ratings were still lower for the latter. The key finding from this experiment is that native speakers of English fail to recognize multiple negation sentences as acceptable even though they are perfectly grammatical. This finding is relevant in several ways.

First, under the *changing encodings hypothesis*, participants are expected to access a fully encoded final-stage interpretation when given ample time. Therefore – in parallel with the pattern observed for NPI illusions – multiple negation sentences were expected to be recognized as acceptable in offline acceptability tasks as final-stage computations are supposed to be available. Instead, there is a clear conflict between grammatical knowledge and offline judgments. The results show an interesting alignment between online and offline responses, and this may question the interpretation of the findings as an illusion of ungrammaticality; at least considering a narrow definition of grammatical illusions. In the general discussion, we will put together the results from the three experiments and examine

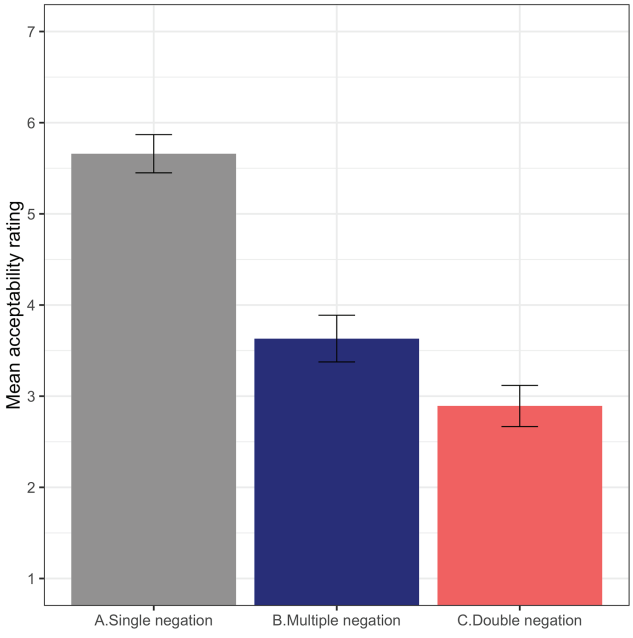


FIGURE 3 | Average acceptability rating for the experimental conditions aggregated by participant (Experiment 3). Error bars indicate standard error of the mean.

what they tell us about parser-grammar misalignments and grammatical illusions. Nonetheless, it is important to note that processing principles alone may not be enough to account for the penalized ratings of multiple negation sentences. Extralinguistic factors related to the stigmatization of negative concord varieties of English and prescriptive bans against the use of double negation (Nevalainen, 2006; Horn, 2010) could have contributed to the surprisingly low ratings attributed to these grammatical sentences; particularly in an experimental design in which they are intermixed with actual double negation sentences. In this context, the mere presence of two negative elements could have guided participants' decisions even when they had unlimited time to provide a response.

Second, the fact that multiple negation sentences are highly penalized in offline ratings provides the strongest case against the *ever-never confusability hypothesis*. The processing effects observed in Experiments 1 and 2 provided initial support against this account. Yet, the different baselines used in this research (single and double negation) and in classic NPI illusions (licensed and unlicensed NPIs) make it difficult to straightforwardly map the online behavior of multiple negation sentences to NPI illusion sentences. Likert scales, the dependent variable in offline ratings, provide a slightly less baseline-conditioned indication of the perceived status of multiple negation sentences. Even though single and double negation sentences act as anchors, the experimental conditions were also intermixed with other grammatical violations that helped participants setting a threshold. Still, multiple negation sentences were given a mean rating of 3.6/7, a score that is on a par with the mean obtained by ungrammatical fillers. These results provide robust evidence that sentences containing *never* instead of *ever* are highly dispreferred by native speakers of English. The fact that they are not be able to recognize them as grammatical is in conflict with the idea that such representations could somehow rescue NPI illusions in processing tasks.

GENERAL DISCUSSION

The series of experiments presented here used online (i.e., speeded judgments and self-paced reading) and offline (i.e., acceptability judgments) methods as a means to study different grammatical configurations of negative elements. The focus of the project was on multiple negation sentences – condition B, repeated in (11) – which displayed the negative markers *no* and *never* in different clauses. The primary objective of this project was studying the online and offline perception of these sentences. To this end, we compared them with similar sentences without the negative element in the relative clause (i.e., single negation, condition A), and with sentences in which both *no* and *never* appeared in the main clause (i.e., double negation, condition C). The observed pattern of results was consistent across experimental measures in showing that multiple negation sentences incurred in an increased processing cost (Experiments 1 and 2) and were also perceived as less acceptable (Experiments 1 and 3) than equivalent single negation sentences. Importantly, the responses for the double negation condition across the three tasks indicate a more degraded perception and slower recovery from disruption.

(11) The authors [that **no** critics recommended] have **never** received an acknowledgement for a best-selling novel.

The fact that double negation sentences were strongly rejected confirms the initial assumptions to conceive them as a degraded baseline. Moreover, including this manipulation in the design was interesting in itself, given the limited attention that the phenomenon of double negation has received in psycholinguistics. Apart from Schiller et al. (2017), who focused on simpler combinations of verbal and affixal negation (e.g., *not unhappy*), this is, to the best of our knowledge, the first psycholinguistic study that uses time-sensitive measures to investigate double negative dependencies. Even though Standard English is commonly classified as a double negation language, this research shows that double negative dependencies do not come at free cost for the language user. This is not surprising considering that the pragmatic function of double negation is to contradict or correct a previous negative statement (Horn, 1991; Puskás, 2012), and thus, double negatives are subject to restricted pragmatic licensing conditions. As described in the introduction, double negatives have been found to appear in specific information structure configurations (Larrivée, 2016) and to be signaled by certain prosodic cues such as contradictory contour (Espinal and Prieto, 2011; Prieto et al., 2013). In addition, this investigation provides evidence that native speakers display strong processing disruptions when double negation dependencies are encountered in isolation. This finding emphasizes the mentioned pragmatic licensing requirements as a condition for double negatives to be interpreted, placing the grammar of double negation at the interface of syntax and pragmatics.

The result that participants consistently reject double negative dependencies overrules one potential concern of this research: the possibility that the participants in the experiments had grammars that allowed negative concord configurations. Native speakers of English are often exposed to instances of negative concord dependencies (e.g., *I cannot get no satisfaction*) as they are allowed in many contemporary varieties of English (e.g., African American Language or Appalachian English). In fact, some theoretical proposals (e.g., Zeijlstra, 2004; Tubau, 2008; Blanchette, 2013, 2015) have hypothesized that the underlying structure of Standard English is that of negative concord. In this vein, Blanchette and Lukyanenko (2019) demonstrate that, in the absence of the necessary licensing conditions, native speakers of English can actually interpret double negation dependencies as negative concord. The participants in our experiments were not explicitly tested for having grammars that allowed negative concord dependencies in order to avoid calling attention to the manipulation. However, we assume that interpreting double negation conditions as a case of negative concord should have facilitated its processing. On the contrary, the results regarding multiple negation and double negation conditions are the opposite to what one would expect if participants' grammars allowed for negative concord structures. Nonetheless, the strong reactions against double

negation were possibly exacerbated by two factors: first, *no* and *never* are not a frequent negative concord or double negation configuration. Second, the participants in the tasks were university educated speakers of English. As Thornton et al. (2016) pointed out, people in academic settings are generally aware of the social stigma associated with negative concord and with prescriptive views on double negation. In sum, the empirical evidence does not support the possibility that participants could be parsing the two negative elements as forming a negative concord dependency.

The main aim of this research was to test two contrasting predictions made for multiple negation sentences on the basis of previous NPI illusion accounts. On the one hand, the *ever-never confusability hypothesis* predicted that these sentences should come across as well-formed in English, and accordingly, they should be processed without problems. The results from the three experiments provide compelling evidence against this hypothesis. On the other hand, based on the *changing encodings hypothesis*, it was predicted that the negative quantifier inside the relative clause could interfere with the integration of *never*, generating an illusion of ungrammaticality. Under this rationale, despite the online interference, it was initially assumed that comprehenders should recognize multiple negation sentences as acceptable when given ample time. Instead, multiple negation sentences are consistently given low ratings in the untimed judgment task, making it less straightforward to map the relation between multiple negation sentences and NPI illusions. The connection between the two phenomena and the possible sources of the degraded perception of grammaticality is explored below.

Relating Multiple Negation Sentences to Negative Polarity Item Illusions

Parker and Phillips' (2016) account of NPI illusions explained spurious licensing as the consequence of accessing incomplete representations of the relative clause material when the NPI is encountered. Their account shifted the attention from the previously proposed erroneous application of NPI-specific licensing mechanisms (i.e., Vasishth et al., 2008; Xiang et al., 2009, 2013) to changes in the encoding of the representations that are used for licensing. In doing so, they provided the basis for an interesting parallelism between NPI illusion sentences and similar sentences containing *never*: if the negative quantifier is accessible to spuriously license the NPI when *ever* is encountered soon after the relative clause, it may also be accessible when *never* is encountered in the same position. The slow RTs observed for multiple negation sentences at the pre-critical region are taken as evidence that at least some aspects of the relative clause material are still being encoded, and thus, that individual feature values – such as negation – could still be transparently accessible. Even though the adverb *never*, as sentential negation, does not need to be licensed by a dependency with any previous element, under a cue-based architecture it assumed that "each incoming words triggers retrievals to integrate that word with the preceding structure" (Lewis et al., 2006, p. 448). If the embedded negation is active when *never* is being integrated, we speculate that the observed difficulties could be indexing the parser's evaluation of a possible dependency between *no* and *never*. Given that double negative dependencies are shown to generate strong processing problems, similar problems are expected to emerge if the parser entertains a relation between *no* and *never* in multiple negation sentences.

The disruptions observed in Experiments 1 and 2 are compatible with this interpretation, and we argue that they could be understood as an illusion of ungrammaticality. Nonetheless, if this phenomenon represents the opposite case of NPI illusions, it may be initially surprising that comprehenders are unable to perceive multiple negation sentences as acceptable in untimed ratings, since they are uncontroversially grammatical. How are the low ratings explained, then? Even though offline judgments are generally conceived as a measure of acceptability, it is widely known that they are sensitive to issues of processability and have been reliably used to uncover processing effects (e.g., Fanselow and Frisch, 2006; Sprouse, 2008; Hofmeister et al., 2013; Dillon et al., 2017). With this in mind, the low ratings for multiple negation sentences could arise from the difficulties integrating *never* in the context of *no*, particularly if a temporary double negative dependency is being temporarily entertained, prompting participants to give low ratings based on simpler heuristics such as the mere presence of two negative elements. In this way, the results from Experiments 1 to 3 are compatible with an interpretation in terms of illusion of ungrammaticality. Yet, there is an alternative – and perhaps simpler – account that deserves exploring: the disruption observed for multiple negation sentences could simply reflect the parser's limitations in processing sentences with two negations.

Integrating a negation is a complex operation that is known to impact the incremental interpretation of sentences. In multiple negation sentences, the parser must undergo this process twice: first inside the relative clause and, then, in the main clause. Processing difficulty, understood as a measure of the resources required to compute the correspondences between forms and meanings (Culicover, 2013), can accumulate during sentence processing in such a way that it can produce additive effects (e.g., Gibson, 1990; Kluender and Kutas, 1993). Thus, one could speculate that the comprehension system may not be able to handle the additive syntactic, semantic and pragmatic complexity of two negative operations when they appear close in the input. In multiple negation sentences, this processing overload is expected to originate when the second negation (i.e., *never*) is encountered if the first negation (i.e., *no*) is still being integrated, exceeding the computational resources of the system. As a consequence, grammar-independent factors related to the limitations of human parser may impede the identification of the correct grammatical analysis, resulting in the processing problems and low acceptability ratings observed. Assuming that processing complexity alone can account for the results eliminates the need to appeal to intermediate stages of representation building and the temporary evaluation of a dependency between *no* and *never* as the source of the effects. In some respects, this interpretation of the findings treats multiple negation sentences on a par with other patterns of misalignment like multiple center embeddings[8]. Indeed, some authors (e.g., Bever, 1970;

[8]Example (4) repeated here: *The patient who the nurse who the clinic had hired met Jack.

Corblin, 1996) have conceptually associated the complexity of negation to that of multiple embedding. Multiple center-embedding sentences reflect the limitations of the parser to generate a representation that is nonetheless available in the grammatical repertoire. Along the same lines, multiple negation sentences could represent another instantiation of the computational limitations of the comprehension system.

If an explanation based solely on processing complexity is the right characterization of the empirical evidence, this limitation of the human parser is expected to extend to similar sentences containing two negative markers. Nonetheless, a number of observations suggest that native speakers of English are able to generate valid representations for sentences that contain two negative elements. For instance, speakers of English can, presumably, understand and express sentences like (12) in spite of their relative complexity.

(12) I did not promise that I would not go.

Sentences like (12) are unsurprising from the perspective of theoretical linguistics because each negative element can only be interpreted independently and, thus, each clause illustrates an instance of single negation. This may explain why these type of constructions are only mentioned in passing by theoretical linguistic works, which describe them as unproblematic and frequent in natural languages (Huddleston and Pullum, 2002; Zeijlstra, 2004). In a recent work using the truth-value judgment task (Crain and Thornton, 1998), Thornton et al. (2016) compared adult and children's interpretation of sentences with double negation and negative concord dependencies. In order to assess the possibility that children could experience problems with two negations simply due to processing limitations, they included sentences like (13) as a control condition.

(13) The girl who did not skip bought nothing.

Similar to our multiple negation sentences, the control condition in Thornton et al. (2016) contained two independent negative markers in different clauses: one inside a relative clause (i.e., *did not*) and the other in the main clause (i.e., *nothing*). If an explanation based on processing is on the right track these sentences are expected to be problematic. However, the results by Thornton et al. (2016) do not seem to point in this direction, as neither adults nor children exhibited problems with them. Importantly, though, the task in Thornton et al. (2016) was a truth-value judgment, which was presented in a context. Although further research should consider this more carefully, the evidence so far suggests that native speakers of English can indeed parse sentences with two negative markers, and thus, that multiple negation sentences and multiple center-embedding should not be conceptualized as analogous cases. Furthermore, there are two remarkable differences between Thornton et al.'s controls and our multiple negation sentences that strengthen the parallelisms with NPI illusions. First, in (13), the main clause negation *nothing* appears after the main clause verb (i.e., *bought*). In our stimuli, *never* appears before the verb, and thus, closer to the relative clause. This is an interesting fact if we take into account that Parker and Phillips' (2016) study demonstrates that illusory licensing disappears when the unlicensed NPI is located after the main clause verb (see example 9). Second, whereas the intervening negation in multiple negation sentences is a negative quantifier, the control sentences by Thornton and colleagues use verbal negation. In a recent investigation, de-Dios-Flores et al. (2017) found that the classic NPI illusion pattern does not occur when the intervening negation is verbal negation, suggesting that NPI illusions arise at least in part as a result of the use of quantificational licensors in the relative clause (cf. Muller et al., 2019).

Considering the above, it is possible that differences in the type and relative position of the negations could explain the contrast between the difficulties generated by multiple negation sentences and the apparent ease with which sentences like (13) are interpreted by both adults and children. These observations about NPI illusions generate interesting predictions for multiple negation sentences. In particular, further research should clarify the role of distance and type of negation in the processing problems observed in multiple negation sentences and also the possible interpretations that speakers ascribe to multiple and double negation sentences. In light of the above, it seems unlikely that the processing problems and degraded perception of multiple negation sentences are solely explained by the additive complexity of integrating two negations. Indeed, if comprehenders were unable to deal with these sentences simply because they have two negations, multiple and double negation sentences should pattern alike in the three tasks. Contrary to this, the differences between these two conditions is patent across tasks and measurements. This is particularly evident in Experiment 1, in which multiple negation sentences were accepted in more than 60% of the cases whereas the acceptance of double negation sentences was below 30%.

The degree of similarity between NPI illusions and multiple negation suggests that the same incomplete encodings that ameliorate the online perception of unlicensed NPIs could be responsible for deteriorating the online perception of grammatical multiple negation sentences. This interpretation of the results generalizes Parker and Phillips' (2016) account of NPI illusions to other configurations, with the additional assumption that the low ratings are the combined product of these processing difficulties and simpler heuristics such as the mere presence of two negative elements. Such heuristics could have been developed by participants as a consequence of the existing social stigmas and prescriptive bans against negative concord and double negation. The hypothesized intrusion of extralinguistic pressures is supported by the fact that multiple negation sentences were actually more penalized when participants had unlimited time (Experiment 3) than when they were asked to provide fast judgments (Experiment 1). By way of conclusion, the next section tries to integrate these findings in the broader context of misalignments.

Widening the Grammatical Illusions Landscape

This research has taken NPI illusions as a starting point in order to examine a candidate structure for a case of illusion of ungrammaticality. To this end, our stimuli were created by substituting the NPI *ever* in Parker and Phillips' (2016) illusion

stimuli by the negative adverb *never*. The results confirm that the integration of the adverb *never* in the main clause is disrupted by the presence of a linearly preceding but structurally inaccessible negative quantifier, resulting in perceived unacceptability of grammatical sentences. The previous section discussed two possible explanations for this interesting pattern of misalignment: one possibility is that they reflect an arbitrary failure of the system due to processing complexity. Another possibility is that the problems attested in multiple negation sentences can be predicted from the same erroneous computations that cause NPI illusions. Our evaluation of the evidence points to the latter, although further research is necessary in order to clarify the degree of similarity between the two phenomena. Either way, multiple negation sentences represent a hitherto unknown case of misalignment that opposes grammatical knowledge with online/offline responses. Conceptualizing it as an illusion of ungrammaticality invites a reflection on the definition and scope of the concept of grammatical illusions.

In the introductory section of this paper, agreement attraction and the spurious licensing of NPIs were presented as paradigmatic examples of grammatical illusions. In this context, the concept of grammatical illusions is generally reserved to describe cases in which grammatical violations do not seem to be perceived in online measures but are then perfectly identified when comprehenders are given ample time. This characterization of grammatical illusions comes with two important assumptions: first, that illusory processes do not affect offline ratings, and second, that comprehenders are not thought to experience the opposite phenomenon (i.e., illusions of ungrammaticality). Even though agreement and NPI illusions often fit into this narrow definition, careful examination of the empirical evidence does not always support such a neat characterization. With regard to the first assumption, previous studies have actually reported an improved perception for NPI illusion sentences also in acceptability judgments, even when the amelioration effects are much weaker than those obtained in online tasks (e.g., Xiang et al., 2006; de-Dios-Flores et al., 2017; Yanilmaz and Drury, 2018)[9]. Thus, there is evidence that grammatical illusions do sometimes affect offline ratings. In addition, comparative illusions (Wellwood et al., 2018) and the presence of agreement attraction effects in production tasks (Bock and Miller, 1991; Bock et al., 2012) are other examples that grammatical illusions do not always surface as neat differences between online and offline responses. With regard to the second assumption, in addition to the evidence collected in this project, there are examples in the literature that could be classified as illusions of ungrammaticality. For instance, in the study of agreement dependencies, some researchers have reported that attraction effects also affect agreement relations in grammatical sentences (e.g., Acuña-Fariña et al., 2014; Lago et al., 2015; Laurinavichyute and von-der-Malsburg, 2019). Along the same lines, it has been shown that the perception of perfectly grammatical unagreement dependencies is degraded in online measures (Mancini et al., 2014; Mancini, 2018).

[9]It is important to note that not every study on NPI illusions reports acceptability ratings for their stimuli (e.g. Vasishth et al., 2008; Xiang et al., 2009).

This varied pattern of misalignments challenges the narrow definition of grammatical illusions because it leaves out many interesting effects, limiting the characterization of the existing evidence and our understanding of the connections among different phenomena – e.g., between multiple negation sentences and NPI illusions. If linguistic illusions are understood as mismatches between grammatical knowledge and the outcomes of language comprehension, a wider illusory space should include misalignments that affect both grammatical and ungrammatical sentences as well as permanent (i.e., online and offline) and temporary (i.e., online) effects. Such a broader conceptualization of illusion-like phenomena would not need to capitalize on black and white distinctions between online and offline responses, while it should still delve deeper on the reasons why different types of dependencies yield different patterns of misalignment in online and offline tasks. Specific linguistic configurations – like multiple negation sentences or NPI illusions – are not ultimately investigated in order to understand them in isolation; but rather, to understand their connections and integrate them into a theory of how misalignments emerge.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was supported by the Spanish Ministry of Education (grant FPU2013/01628), the Spanish Ministry of Economy and Competitiveness (grant PSI2015-65116-P), and the Autonomous Government of Galicia (grant ED431B-2019/2021).

ACKNOWLEDGMENTS

The author wishes to thank Carlos Acuña-Fariña, Colin Phillips, Hanna Muller, and the audiences at BICLCE-2017, AMLAP-2018, and AEDEAN-2018 for helpful discussions during earlier stages of the project, and two reviewers for providing insightful

comments on the work. Furthermore, the author would like to express her gratitude to Dan Parker and Colin Phillips for their generosity in allowing her to modify their original experimental items and to the Cognitive Neuroscience of Language Laboratory of the University of Maryland for providing the technical resources necessary to collect the data.

REFERENCES

Acuña-Fariña, J. C., Meseguer, E., and Carreiras, M. (2014). Gender and number agreement in comprehension in Spanish. *Lingua* 143, 108–128. doi: 10.1016/j.lingua.2014.01.013

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Barker, C. (1970). Double negatives. *Linguist. Inquiry* 1, 169–186.

Barker, C. (2018). Negative polarity as scope marking. *Linguist. Philos.* 41, 483–510. doi: 10.1007/s10988-018-9234-2

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bever, T. G. (1970). "The cognitive basis for linguistic structures" in *Cognition and language development*. ed. J. R. Hayes (New York, EEUU: Wiley & Sons, Inc.), 277–360.

Blanchette, F. (2013). Negative concord in English. *Linguist. Var.* 13, 1–47. doi: 10.1075/lv.13.1.01bla

Blanchette, F. (2015). English negative concord, negative polarity, and double negation. PhD dissertation. City University of New York.

Blanchette, F., and Lukyanenko, C. (2019). Unacceptable grammars? An eye-tracking study of English negative concord. *Lang. Cogn.* 11, 1–40. doi: 10.1017/langcog.2019.4

Blanchette, F., Nadeu, M., Yeaton, J., and Deprez, V. (2018). English negative concord and double negation: the division of labor between syntax and pragmatics. *Proc. Linguist. Soc. Am.* 3, 1–15. doi: 10.3765/plsa.v3i1.4349

Bock, K., Carreiras, M., and Meseguer, E. (2012). Number meaning and number grammar in English and Spanish. *J. Mem. Lang.* 66, 17–37. doi: 10.1016/j.jml.2011.07.008

Bock, K., and Miller, C. A. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93. doi: 10.1016/0010-0285(91)90003-7

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1999). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *NeuroImage* 10, 216–224. doi: 10.1006/nimg.1999.0465

Chomsky, N. (1957). *Syntactic structures*. Oxford, England: Mouton de Gruyter.

Chomsky, N., and Miller, G. A. (1963). "Introduction to the formal analysis of natural languages" in *Handbook of mathematical psychology*. eds. R. D. Luce, R. Bush, and E. Galanter (New York, EEUU: Wiley), 269–321.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd Edn. Hillsdale (EEUU): Lawrence Erlbaum.

Corblin, F. (1996). Multiple negation processing in natural language. *Theoria* 62, 214–259. doi: 10.1111/j.1755-2567.1996.tb00503.x

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgements*. Thousand Oaks, USA: Sage Publications.

Crain, S., and Thornton, R. (1998). *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. Boston, EEUU: MIT Press.

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8, 1–18. doi: 10.1371/journal.pone.0057410

Culicover, P. (2013). *Grammar & complexity: Language at the intersection of competence and performance*. Oxford, England: Oxford University Press.

de Swart, H. (2010). *Expression and interpretation of negation: An OT typology*. The Netherlands: Springer.

de-Dios-Flores, I., Muller, H., and Phillips, C. (2017). Negative polarity illusions: licensors that don't cause illusions, and blockers that do. Poster presented at CUNY, Boston.

Dillon, B., Clifton, C., Sloggett, S., and Frazier, L. (2017). Appositives and their aftermath: interference depends on at-issue vs. not-at-issue status. *J. Mem. Lang.* 96, 93–109. doi: 10.1016/j.jml.2017.04.008

Drenhaus, H., Saddy, D., and Frisch, S. (2005). "Processing negative polarity items. When negation comes through the backdoor" in *Linguistic evidence: Empirical, theoretical, and computational perspectives*. eds. S. Kepser and M. Reis (Berlin: Mouton de Gruyter), 145–165.

Drummond, A. (2013). Ibex farm. Available at: http://spellout.net/ibexfarm

Embick, D., and Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: correlational, integrated and explanatory neurolinguistics. *Lang. Cogn. Neurosci.* 30, 357–366. doi: 10.1080/23273798.2014.980750

Enochson, K., and Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon mechanical Turk. *PLoS One* 10, 1–17. doi: 10.1371/journal.pone.0116946

Espinal, M. T., and Prieto, P. (2011). Intonational encoding of double negation in Catalan. *J. Pragmat.* 43, 2392–2410. doi: 10.1016/j.pragma.2011.03.002

Fanselow, G., and Frisch, S. (2006). "Effects of processing difficulty on judgements of acceptability" in *Gradience in grammar: Generative perspectives*. eds. G. Fanselow, C. Féry, M. Schlesewsky, and R. Vogel (New York, EEUU: Oxford University Press), 291–316.

Ferreira, F., Bailey, K. G. D., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* 11, 11–15. doi: 10.1111/1467-8721.00158

Ferreira, F., and Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Lang Ling Compass* 1, 71–83. doi: 10.1111/j.1749-818X.2007.00007.x

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., and Perry, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology* 20, 400–409. doi: 10.1111/j.1469-8986.1983.tb00920.x

Frank, S. L., Bod, R., and Christiansen, M. H. (2012). How hierarchical is language use? *Proc. R. Soc. B Biol. Sci.* 279, 4522–4531. doi: 10.1098/rspb.2012.1741

Frazier, L. (1985). "Syntactic complexity" in *Natural language processing: Psychological, computational and theoretical perspectives*. eds. D. Dowty, L. Karttunen, and A. Zwiky (Cambridge, UK: Cambridge University Press), 129–189.

Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* 14, 178–210. doi: 10.1016/0010-0285(82)90008-1

Gelman, A., and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models* (*analytical methods for social research*). Cambridge, UK: Cambridge University Press.

Giannakidou, A. (1998). *Polarity sensitivity as (non)veridical dependency*. Amsterdam: John Benjamins.

Giannakidou, A. (2006). Only, emotive factive verbs, and the dual nature of polarity dependency. *Language* 82, 575–603. doi: 10.1353/lan.2006.0136

Giannakidou, A. (2011). "Positive polarity items and negative polarity items: variation, licensing, and compositionality" in *Semantics: An international handbook of natural language meaning*. eds. K. von-Heusinger, C. Maienborn, and P. Portner (Berlin: Mouton de Gruyter), 1660–1712.

Gibson, E. (1990). "A computational theory of processing overload and garden-path effects" in *Proceedings of the 13th conference on computational linguistics*. ed. H. Karlgren (Stroudsburg, PA: Association for Computational Linguistics), 114–119.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., and Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychol. Sci.* 24, 1079–1088. doi: 10.1177/0956797612463705

Gibson, E., Piantadosi, S., and Fedorenko, K. (2011). Using mechanical Turk to obtain and analyze English acceptability judgments. *Lang Ling Compass* 5, 509–524. doi: 10.1111/j.1749-818X.2011.00295.x

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cogn. Process.* 14, 225–248. doi: 10.1080/016909699386293

Gimenes, M., Rigalleau, F., and Gaonach, D. (2009). When a missing verb makes a French sentence more acceptable. *Lang. Cogn. Process.* 24, 440–449. doi: 10.1080/01690960802193670

Häussler, J., and Bader, M. (2015). An interference account of the missing-VP effect. *Front. Psychol.* 6, 1–16. doi: 10.3389/fpsyg.2015.00766

Herbert, C., and Kübler, A. (2011). Dogs cannot bark: event-related brain responses to true and false negated statements as indicators of higher-order conscious processing. *PLoS One* 6:e25574. doi: 10.1371/journal.pone.0025574

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Lang. Cogn. Process.* 26, 376–405. doi: 10.1080/01690965.2010.492642

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., and Snider, N. (2013). The source ambiguity problem: distinguishing the effects of grammar and processing on acceptability judgments. *Lang. Cogn. Process.* 28, 48–87. doi: 10.1080/01690965.2011.572401

Horn, L. R. (1991). "Duplex negation affirmat ...: the economy of double negation" in *Papers from the 27th regional meeting of the Chicago linguistic society. Part two: The parasession on negation*. eds. L. M. Dobrin, L. Nichols, and R. M. Rodriguez (Chicago, EEUU: Chicago Linguistic Society), 78–106.

Horn, L. R. (2001). *A natural history of negation*. Stanford: CSLI.

Horn, L. R. (2010). "Multiple negation in English and other languages" in *The expression of negation* (Berlin, Boston: De Gruyter Mouton), 111–148.

Huddleston, R., and Pullum, J. (2002). *Cambridge grammar English language | grammar and syntax*. Cambridge, UK: Cambridge University Press.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

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

Kadmon, N., and Landman, F. (1993). Any. *Linguist. Philos.* 16, 353–422. doi: 10.1007/BF00985272

Kaup, B., Lüdtke, J., and Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: is a door that is not open mentally closed? *J. Pragmat.* 38, 1033–1050. doi: 10.1016/j.pragma.2005.09.012

Kluender, R., and Kutas, M. (1993). Subjacency as a processing phenomenon. *Lang. Cogn. Process.* 8, 573–633. doi: 10.1080/01690969308407588

Krifka, M. (1995). The semantics and pragmatics of polarity items. *Linguist. Anal.* 25, 209–257.

Ladusaw, W. A. (1979). Negative polarity items as inherent scope relations. PhD dissertation. University of Texas.

Lago, S., Shalóm, D. E., Sigman, M., Lau, E. F., and Phillips, C. (2015). Agreement attraction in Spanish comprehension. *J. Mem. Lang.* 82, 133–149. doi: 10.1016/j.jml.2015.02.002

Laka, I. (1994). *On the syntax of negation*. New York, London: Garland Publishing.

Larrivée, P. (2016). "The markedness of double negation" in *Negation and polarity: Experimental perspectives language, cognition, and mind*. eds. P. Larrivée and C. Lee (Berlin, Germany: Springer International Publishing), 177–198.

Laurinavichyute, A., and von-der-Malsburg, T. (2019). Agreement attraction effects in the comprehension of grammatical sentences. Poster presented at CUNY, Boulder.

Levy, R. (2008a). "A noisy-channel model of rational human sentence comprehension under uncertain input" in *EMNLP '08: Proceedings of the conference on empirical methods in natural language processing*. eds. M. Lapata and H. T. Ng (Stroudsburg, PA: Association for Computational Linguistics), 234–243.

Levy, R. (2008b). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R., Bicknell, K., Slattery, T., and Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *PNAS* 106, 21086–21090. doi: 10.1073/pnas.0907664106

Lewis, S., and Phillips, C. (2015). Aligning grammatical theories and language processing models. *J. Psycholinguist. Res.* 44, 27–46. doi: 10.1007/s10936-014-9329-z

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Linebarger, M. C. (1987). Negative polarity and grammatical representation. *Linguist. Philos.* 10, 325–387. doi: 10.1007/BF00584131

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703. doi: 10.1037/0033-295X.101.4.676

Mancini, S. (2018). When grammar and parsing agree. *Front. Psychol.* 9, 1–5. doi: 10.3389/fpsyg.2018.00336

Mancini, S., Molinaro, N., Davidson, D. J., Avilés, A., and Carreiras, M. (2014). Person and the syntax-discourse interface: an eye-tracking study of agreement. *J. Mem. Lang.* 76, 141–157. doi: 10.1016/j.jml.2014.06.010

Miller, G. A., and Isard, S. (1964). Free recall of self-embedded English sentences. *Inf. Control.* 7, 292–303. doi: 10.1016/S0019-9958(64)90310-9

Millisecond Software (2015). Inquisit 4 computer software. Available at: https://www.millisecond.com

Muller, H., de-Dios-Flores, I., and Phillips, C. (2019). Not (just) any licensors cause negative polarity illusions. Poster presentation at CUNY, Boulder.

Nevalainen, T. (2006). Negative concord as an English "vernacular universal": social history and linguistic typology. *J. Engl. Linguist.* 34, 257–278. doi: 10.1177/0075424206293144

Nicenboim, B., Logačev, P., Gattei, C., and Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: working memory and locality effects. *Front. Psychol.* 7:280. doi: 10.3389/fpsyg.2016.00280

Parker, D., and Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition* 157, 321–339. doi: 10.1016/j.cognition.2016.08.016

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456. doi: 10.1006/jmla.1999.2653

Phillips, C., and Lewis, S. N. (2013). Derivational order in syntax: evidence and architectural consequences. *Stud. Linguist.* 6, 11–47.

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension" in *Experiments at the interfaces syntax and semantics*. ed. J. Runner (Leiden, The Netherlands: Brill), 147–180.

Prieto, P., Borràs-Comes, J., Tubau, S., and Espinal, M. T. (2013). Prosody and gesture constrain the interpretation of double negation. *Lingua* 131, 136–150. doi: 10.1016/j.lingua.2013.02.008

Pullum, J. K. (2004). Plausible angloid gibberish. Language log. Available at: http://itre.cis.upenn.edu/~myl/languagelog/archives/000860.html (Accessed May 17, 2019).

Puskás, G. (2012). Licensing double negation in NC and non-NC languages. *Nat. Lang. Linguist. Theory* 30, 611–649. doi: 10.1007/s11049-011-9163-z

R Development Core Team (2014). R: a language and environment for statistical computing. Available at: http://www.r-project.org/

Schiller, N. O., van Lenteren, L., Witteman, J., Ouwehand, K., Band, G. P. H., and Verhagen, A. (2017). Solving the problem of double negation is not impossible: electrophysiological evidence for the cohesive function of sentential negation. *Lang. Cogn. Neurosci.* 32, 147–157. doi: 10.1080/23273798.2016.1236977

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *J. Verbal Learn. Verbal Behav.* 15, 143–157. doi: 10.1016/0022-5371(76)90015-3

Sprouse, J. (2008). The differential sensitivity of acceptability judgments to processing effects. *Linguist. Inquiry* 39, 686–694. doi: 10.1162/ling.2008.39.4.686

Sprouse, J. (2011). A validation of Amazon mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behav. Res. Methods* 43, 155–167. doi: 10.3758/s13428-010-0039-7

Sprouse, J., and Lau, E. F. (2013). "Syntax and the brain," in *The Cambridge handbook of generative syntax*. ed. M. den Dikken (Cambridge: Cambridge University Press), 971–1005.

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua* 134, 219–248. doi: 10.1016/j.lingua.2013.07.002

Staub, A. (2009). On the interpretation of the number attraction effect: response time evidence. *J. Mem. Lang.* 60, 308–327. doi: 10.1016/j.jml.2008.11.002

Thornton, R., Notley, A., Moscati, V., and Crain, S. (2016). Two negations for the price of one. *Glossa J. Gen. Linguist* 1, 1–30. doi: 10.5334/gjgl.4

Townsend, D., and Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Boston, EEUU: MIT Press.

Trotzke, A., Bader, M., and Frazier, L. (2013). Third factors and the performance interface in language design. *Biolinguistics* 7, 1–34.

Tubau, S. (2008). *Negative concord in English and romance: Syntax-morphology interface conditions on the expression of negation*. Amsterdam, Holland: LOT Publications.

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Wason, P. C. (1961). Response to affirmative and negative binary statements. *Br. J. Psychol.* 52, 133–142. doi: 10.1111/j.2044-8295.1961.tb00775.x

Wellwood, A., Pancheva, R., Hacquard, V., and Phillips, C. (2018). The anatomy of a comparative illusion. *J. Semant.* 35, 543–583. doi: 10.1093/jos/ffy014

Xiang, M., Dillon, B., and Phillips, C. (2006). *Testing the strength of the spurious licensing effect for negative polarity items*. New York: Talk at CUNY.

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xiang, M., Grove, J., and Giannakidou, A. (2013). Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Front. Psychol.* 4, 1–19. doi: 10.3389/fpsyg.2013.00708

Yanilmaz, A., and Drury, J. E. (2018). Prospective NPI licensing and intrusion in Turkish. *Lang. Cogn. Neurosci.* 33, 111–138. doi: 10.1080/23273798.2017.1371779

Yun, J., Lee, S. Y., and Drury, J. E. (2018). "Negative polarity illusion in Korean" in *Proceedings of WAFL*. eds. C. Guillemot, T. Yoshida, and S. J. Lee (Cambridge, USA: MIT Press).

Zeijlstra, H. (2004). Sentential negation and negative concord. PhD dissertation. Utrecht, Holland: University of Amsterdam.

Zeijlstra, H. (2007). Negation in natural language: on the form and meaning of negative elements. *Lang Ling Compass* 1, 498–518. doi: 10.1111/j.1749-818X.2007.00027.x

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 12 November 2019
doi: 10.3389/fpsyg.2019.02486

Asymmetries in the Acceptability and Felicity of English Negative Dependencies: Where Negative Concord and Negative Polarity (Do Not) Overlap

Frances Blanchette[1] and Cynthia Lukyanenko[2]*

[1] Center for Language Science and Department of Psychology, Penn State University (PSU), University Park, TX, United States, [2] Linguistics Program, Department of English, George Mason University, Fairfax, VA, United States

OPEN ACCESS

Edited by:
M. Teresa Espinal, Autonomous University of Barcelona, Spain

Reviewed by:
Rosalind Jean Thornton, Macquarie University, Australia
Jacee Cho, University of Wisconsin-Madison, United States

***Correspondence:**
Frances Blanchette
fkb1@psu.edu

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

Received: *02 August 2019*
Accepted: *22 October 2019*
Published: *12 November 2019*

Citation:
Blanchette F and Lukyanenko C (2019) Asymmetries in the Acceptability and Felicity of English Negative Dependencies: Where Negative Concord and Negative Polarity (Do Not) Overlap. Front. Psychol. 10:2486. doi: 10.3389/fpsyg.2019.02486

Negative Concord (NC) constructions such as *the news anchor didn't warn nobody about the floods* (meaning "the news anchor warned nobody"), in which two syntactic negations contribute a single semantic one, are stigmatized in English, while their Negative Polarity Item (NPI) variants, such as *the news anchor didn't warn anybody about the floods*, are prescriptively correct. Because acceptability is often equated with grammaticality, this pattern has led linguists to treat NC as ungrammatical in "Standard" or standardized English (SE). However, it is possible that SE grammars do generate NC sentences, and their low incidence and acceptability is instead due to social factors. To explore this question, and the relationship between NC and NPI constructions, we compared the acceptability of overtly negative noun phrases (e.g., *nobody*), NPIs (e.g., *anybody)*, and bare plurals (e.g., *people*), in negative contexts and in conditionals. Negative items were followed by a consequence which supported their single negative meaning, while conditional items were followed by a consequence compatible with the NPI and the bare plural but not the negative noun phrase. Acceptability ratings of the critical NC sentences were reliably lower than constructions with NPIs and bare plurals, but the consequences for all three of these sentence types were rated highly. This reflects an asymmetry in participants' acceptance of NC and their readiness to interpret it in context. A follow-up study with only conditionals revealed that speakers can also find NPIs infelicitous in conditional contexts with consequences that are compatible with a negative interpretation of the NPI, and that negative arguments are felicitous in these same contexts. Taken together, the results support the hypothesis that speakers who do not accept NC have grammars that generate both NC and NPI constructions, and further, that these speakers have two underlying structures for *any*-NPIs in English.

Keywords: acceptability, conditionals, experimental approaches, felicity, grammaticality, Negative Concord, Negative Polarity

INTRODUCTION

Human languages display diversity in whether and how they instantiate negative dependencies (Auwera and Alsenoy, 2016). In a subset of languages, negative arguments are typically found in Negative Concord (NC) constructions, in which two or more syntactic negations contribute a single semantic negation, as in the following Italian example from Zanuttini (1997, p. 8, ex. (13a)):

(1) *(*Non*) *ho visto nessuno*.
NEG have seen nobody
'It is not the case that I have seen somebody.'

In (1), the preverbal negative marker *non* and the negative direct object argument *nessuno* "nobody" are interpreted as a single semantic negation, reflecting a pattern typical to NC constructions[1].

Other languages instantiate negative dependencies through Negative Polarity Item (NPI) constructions[2]. These NPI constructions are similar to the NC construction seen in (1), but they do not have an overtly negative noun phrase. Instead, in place of a phrase like *nessuno* "nobody" in (1), they contain a phrase which is not overtly negative but depends on a preceding element, prototypically a negation, for its licensing. The following is example is from Ewe (Collins et al., 2017, p. 2, ex. (2b)):

(2) Kofi *(mé)-kpɔ́ ame ádéké.
Kofi NEG-see person any
'Kofi didn't see anybody.'

In (2), the term *ádéké* "any" is an NPI. It is not overtly marked for negation, but the negative marker *mé* is required for acceptability, in a manner similar to NC in Italian and other languages.

NC constructions are often modeled as a syntactic dependency between negative elements within a clause (e.g., Haegeman and Zanuttini, 1996; Zeijlstra, 2004; Déprez, 2011; Puskás, 2012; Blanchette, 2013). This is due to the requirement that the preverbal marker be present in the structure as in (1), in conjunction with the resumptive morphological marking of negation[3]. The grammatical nature of NPI constructions is subject to debate, but since Ladusaw (1979) a common view is that they primarily reflect a semantic-pragmatic dependency between the NPI and its licensing context [the negation in (2); e.g., Krifka, 1995; Giannakidou, 1998, 1999, 2002; Zwarts, 1998; Gajewski, 2011; Chierchia, 2013].

English is among the languages which instantiate both NC and NPI constructions. In vernacular English varieties, spontaneous speech reflects variation in negative contexts between these two structure types, as in the following examples from Tortora et al. (2017) *The Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCAppE)[4]. (See Childs, 2017 for an analysis of this type of variation in British vernacular speech corpora.)

(3) They wasn't a radio, they wasn't anything.
'There wasn't a radio, there wasn't anything.'
(AAPCAppE: ALC-FJ-733-1, 0.343)

(4) They wasn't nothing for them to get into.
'There wasn't anything for them to get into.'
(AAPCAppE: ALC-FJ-733-1, 0.478)

Speakers may even employ both construction types within a single utterance, as in the following example from an Appalachian English speaker (cited in Blanchette, 2016, p. 110):

(5) I didn't have no lice, and I didn't have any itch.
'It is not the case that I had lice, and it is not the case that I had itch.'
(AAPCAppE: SKCTC-EA-1, 0.63)

An important and distinguishing feature of English NC is its heavy social stigma (Horn, 2010), a stigma which is not present in other languages with NC. NC is often condemned as illogical, and "Standard" or standardized English (SE) speakers tend to avoid it in usage. Many linguists have taken its unacceptability and absence from SE usage to reflect its underlying ungrammaticality[5]. This is at least in part due to the traditional causal link assumed by linguists between acceptability and grammaticality on the one hand, and unacceptability and ungrammaticality on the other (Etxeberria et al., 2018, p. 2). If there exists a direct connection between acceptability and grammaticality, then it follows that SE grammars generate (prescriptively correct) NPI constructions, but they do not generate NC. Following this line of reasoning further leads to a hypothesis in which utterances such as (5) reflect a form of code-switching between two different grammatical systems. The Appalachian speaker controls two systems, and the component of her grammar that generates the NPI construction overlaps with SE grammars, while the component of her grammar that generates NC does not.

This paper uses experimental means to explore an alternative hypothesis, one which does not assume a direct and causal link between NC unacceptability and ungrammaticality (Lewis and Phillips, 2015; Etxeberria et al., 2018). We acknowledge the social forces shaping NC acceptability, and use measures of meaning in context to contribute toward our understanding of its grammaticality in relation to NPI constructions. We exploit the fact that NPI constructions appear in a broader range of contexts than NC, to illustrate how speakers who do not accept NC nevertheless demonstrate knowledge of

[1] Typical of Romance languages like Catalan, French, Italian, Spanish, and Brazilian Portuguese (e.g., Zanuttini, 1997; Déprez, 2000; Herburger, 2001; De Swart and Sag, 2002; Prieto et al., 2013; Agostini and Schwenter, 2018), NC can also be found many other languages including Afrikaans (Biberauer and Zeijlstra, 2012), Bavarian (Bayer, 1990), Hungarian (Puskás, 2012), West Flemish (Haegeman and Zanuttini, 1996), Serbian/Croation (Progovac, 1994), and others.

[2] See Auwera and Alsenoy (2016, p. 483) on the frequency of negative structure types across languages.

[3] See Giannakidou (2000) for a semantic account, discussed further below.

[4] Following AAPCAppE citation conventions, tokens are followed by the corpus, subcollection, and speaker initials, along with a numerical token identifier.

[5] We follow Hudley and Mallinson (2010) in employing the term standardized English, as opposed to the more common "Standard" English, to acknowledge the agency of prescriptive forces in the standardization process, which excludes variants not because of their lack of systematicity, but rather, because of the identities of the speaker groups who do and do not use them.

when these constructions do and do not overlap in meaning with NPI constructions. We discuss how the results can be taken to support a theory of grammar in which utterances as in (5) do not reflect code-switching, but rather, a form of shifting between surface forms which reflect similar underlying grammatical mechanisms.

ENGLISH NEGATIVE CONCORD AND NEGATIVE POLARITY

This section summarizes several relevant grammatical theories and experimental and psycholinguistic studies of NC and NPI constructions. The literature is vast, and we focus on those most relevant to our experiments. We begin with the assumption that grammars are "abstract descriptions of the representations built by the cognitive system" during language comprehension and production, rather than cognitively real, static references queried by the parser (Lewis and Phillips, 2015, p. 30). Social forces such as prescriptive pressure are external to cognitive representations, but they interact in crucial ways with the outputs of those representations. This is most relevant to studies of NC, which we summarize first.

Negative Concord

The Syntactic Agree Approach

Many recent theories of NC model it as a syntactic Agree relation between negative elements within a clause (e.g., Zeijlstra, 2004; Puskás, 2012; Wallage, 2012; Espinal and Tubau, 2016; Tubau, 2016). Such theories are often motivated, at least in part, by the contrast between NC and so-called Double Negation (DN) constructions, in which each of two syntactic negations contributes a semantic negation. The following examples illustrate:

(6) DN
Speaker A: You're hungry because you ate nothing for lunch.
Speaker B: I didn't eat nothing. I had half a sandwich.
DN meaning: It is not the case that I ate nothing.

(7) NC
Speaker A: I'm hungry.
Speaker B: Me too. I didn't eat nothing.
NC meaning: It is not the case that I ate (something).

Zeijlstra (2004) proposes that sentences such as those in (6) and (7) instantiate two different grammatical systems. Déprez (2011) proposes instead that the distinction is more of a "micro-parametric" one, in which grammars may generate either NC or DN, depending on the syntactic configuration. This "micro-parametric" view is supported by recent experimental work, which has shown that in English as well as in Romance languages, DN constructions as in (6) exist alongside NC constructions as in (7), with DN being reliably associated with a marked prosodic tune relative to the single negation interpretation of NC (Espinal and Prieto, 2011; Espinal et al., 2016; Blanchette et al., 2018; Blanchette and Nadeu, 2018; Déprez and Yeaton, 2018).

Syntactic Agree approaches to modeling NC posit that negative elements are lexically endowed with an uninterpretable feature which needs to be checked in the syntax. Under an Agree approach, the NC sentence in (6) would be modeled roughly as follows (cf. Zeijlstra, 2004):

(8) I did [$_{\text{NegP}}$ ¬[*i*NEG] [$_{\text{Neg}'}$ n't$_{[\sout{u}\text{NEG}]}$ [$_{\text{vP}}$ eat nothing$_{[u\text{NEG}]}$]]]

Example (8) shows how the negative noun phrase *nothing* and the marker *n't* enter the structure with an uninterpretable negative feature [*u*NEG]. By virtue of being uninterpretable, these features must check themselves against the interpretable negative feature [*i*NEG] residing on a phonologically null operator in the head of a higher negative phrase (NegP). This checking relation establishes a syntactic dependency between the semantically non-negative elements *n't* and *nothing* and the semantically negative null operator, yielding an NC structure with a single negative interpretation.

Tubau (2016) represents a recent Agree approach to modeling English NC. She notes that in British English dialects, negative noun phrases need not always be preceded by another negation, and shows how the following variant types are attested:

(9) I didn't eat nothing. (NC)
(10) I ate nothing.

To explain the variation seen in (9) and (10), Tubau proposes a theory in which negative noun phrases such as *nothing* have two distinct lexical entries. The *nothing* in (9) is endowed with an uninterpretable [*u*NEG] feature, which triggers the concord (Agree) relation (as in (8) above), while the *nothing* in (10) has an interpretable [*i*NEG] feature, and thus contributes its own semantic negation without needing to establish an Agree relation with a preceding negative operator. Vernacular British English dialects differ from SE in this theory. In general, SE is assumed to be a DN language, having neither [*i*NEG] nor [*u*NEG]. Instead, each syntactic negation is taken to instantiate an underlying negative operator which is not featurally active and therefore never eligible for Agree, meaning that structures like (9) are not generated.

Negative Concord in Standardized ("Standard") English

While vernacular English varieties are known for instantiating NC (Wolfram and Fasold, 1974; Nevalainen, 2006), a series of recent experimental studies show that SE speakers also have reliable intuitions about this construction type. The studies show that SE speakers have a clear knowledge of the syntactic distribution of NC (Blanchette, 2017), an understanding of its meaning and prosodic properties in relation to DN (Blanchette et al., 2018), and an apparent proclivity toward building NC structures during online processing (Blanchette and Lukyanenko, 2019). These studies all involve comparison of sentences with a negative noun phrase in direct object position following a negative marker as in (11) (and (6/7) above), and sentences with

a negative noun phrase in canonical subject position preceding a negative marker, as in (12)[6,7]:

(11) I didn't see nobody.
(12) Nobody didn't see me.

The results of all three studies demonstrate that SE speakers reliably prefer NC interpretations for sentences like (11), but DN interpretations for sentences like (12).

The sentences in (11) and (12) illustrate a typological divide between what Giannakidou (1998) categorizes as "non-strict" and "strict" NC. Both strict and non-strict NC languages have sentences like (11), in which a negative noun phrase is preceded by and acts in concord with a negated auxiliary, but only strict NC languages have sentences like (12), in which the negative noun phrase both precedes and acts in concord with the negated auxiliary. On the basis of their findings, Blanchette and Lukyanenko (2019, p. 24) therefore suggest that SE may be categorized as "non-strict[8]." They further note a similarity between speakers' subtle intuitions about NC in SE, and more obvious intuitions about parallel NPI constructions. To illustrate, consider the following contrast:

(13) I didn't see anybody.
(14) *Anybody didn't see me[9].

Example (14) shows that NPIs are unacceptable in canonical subject position. Note that (13), which is acceptable, is equivalent in meaning and nearly identical in form to (11), while unacceptable (14) is nearly identical in form to (12)[10]. The acceptability of NPI constructions thus parallels speakers' intuitions about NC, suggesting a possible grammatical relationship between these two construction types. The studies we report in this paper take a first step toward understanding the nature of this relationship, and how it might inform abstract grammatical as well as cognitive theories. To illustrate this, we next provide some background on NPI constructions.

[6]In a study that compares children and adults, Thornton et al. (2016) find that Australian English-speaking adults reliably prefer DN readings over NC readings, in contrast with children, who reliably prefer NC. Children's judgments were elicited in spoken conversation with a puppet, while adults judgments were collected in written form, which suggests that the comparison is not entirely valid. See Blanchette and Lukyanenko (2019) for further discussion of this.

[7]Blanchette (2017) also examines Negative Auxiliary Inversion (NAI) constructions such as the following, in which the negative noun phrase is also in a subject position (for more on NAI see, e.g., Weldon, 1994; Labov, 1972; Green, 2014):

(i) Didn't nobody see me.
'Nobody saw me.'

The specifics are beyond the scope of this paper to discuss, but the general conclusion is that SE speakers prefer NC interpretations for sentences in which the negated auxiliary precedes (and c-commands) the negative noun phrase.

[8]Vernacular Englishes also optionally instantiate the "strict" NC pattern (Labov, 1972; Wolfram and Christian, 1976; Tubau, 2016; among others).

[9]Henry (1995) notes that constructions such as this are possible in Belfast English.

[10]Sentence (14) is argued to be unacceptable because it does not meet the c-command requirement for NPIs and their licensors. We discuss this further below.

Negative Polarity

Downward Entailingness

Ladusaw (1979) observed that NPIs are acceptable when they occur in the scope of a downward entailing expression, which creates "a semantic context which makes inferences run downward on a scale" (p. 179)[11]. The following examples illustrate that negation is downward entailing:

(15) Maria didn't drive.
(16) Maria didn't drive fast.
(17) Maria didn't drive fast and furiously.

The sets denoted by the predicate narrow from (15) to (17), and the entailments hold in that downward direction: If Maria did not drive (the widest set), then it must also be true that she did not drive fast (a narrower set), and that she did not drive fast and furiously (the narrowest set). Note that removing the negation voids this entailment pattern:

(18) Maria drove.
(19) Maria drove fast.
(20) Maria drove fast and furiously.

It can be true that Maria drove, but that she drove slowly and cautiously, which means that (18) being true does not entail that (19) and (20) are also true.

Negation's ability to trigger downward entailing inferences, Ladusaw proposes, is the property that allows it to license NPIs, its removal leading to unacceptability:

(21) Mary didn't drive any cars.
(22) * Mary drove any cars[12].

In addition to this semantic specification, there is also thought to be a syntactic requirement that the NPI be c-commanded by its licensor (Baker, 1970, as cited in Linebarger, 1987, p. 330). Sentence (14), in which the NPI precedes the negation, is one example of why the c-command requirement is needed, since an eligible licensor is present in the structure, but the sentence is nevertheless unacceptable.

Further research on downward entailment for NPI licensing has revealed a number of apparent exceptions to the pattern, one of which is conditionals, which we employ in our experiment. The lack of straightforward downward entailingness in these contexts has led semanticists to expand, refine, or propose alternatives to this as a licensing condition (Giannakidou, 1998, 1999; Von Fintel, 1999; Gajewski, 2011; Chierchia, 2013).

Some recent psycholinguistic studies of NPIs have assumed the downward entailingness theory of NPI licensing in examining speakers' processing of NPIs. Both Vasishth et al. (2008) and Parker and Phillips (2016), for example, investigate so-called

[11]The term "downward entailing" is used synonymously with "monotone decreasing" (Barwise and Cooper, 1981).

[12]For reasons of space and lack of immediate relevance we set aside here and throughout instances of "free choice *any*," as in the following example:

(i) Maria drove any car she wanted.

For a semantic account of free choice *any* (see e.g., Dayal, 1995). For a syntactic account (see Collins and Postal, 2014, p. 43).

"NPI illusions" in which speakers accept and successfully process NPIs despite their not being in the c-command domain of a preceding downward entailing licensor. Interestingly, however, when Szabolsci et al. (2008) set out to confirm via experimental means that NPIs trigger the validation of downward entailing inferences, they found no evidence of a connection between NPI processing and the process of inference validation. This suggests that, while the downward entailingness generalization captures a wide range of facts concerning NPI distribution, it might not be justified after all to assume that this generalization finds a parallel within the actual cognitive mechanisms involved in NPI processing.

A Unified Semantic Theory of NPI and NC Constructions

Giannakidou (1999, 2000) provides an alternative semantic account to explain NPI licensing behaviors, and relates them directly to NC. Under her proposal, "NC is nothing more than a subcase of negative polarity" (Giannakidou, 2000, p. 463)[13]. She argues that noun phrases which participate in NC in Greek (a "strict" NC language) are non-negative universal quantifiers that, like NPIs, are sensitive to the veridicality of their surrounding context[14]. Under her theory, these quantifiers must raise to take scope over a sentential negation. The following is an example of NC in Greek, and the corresponding structure (example (23) is her p. 499 ex. (83), and (24) is adapted from p. 500 ex. (90)):

(23) Dhen ipe o Pavlos TIPOTA[15].
NEG said the Paul n-thing
'Paul said nothing.'

(24) $[_{XP}$ [*tipota*]$_1$ *dhen* $[_{VP}$ *ipe o Pavlos* $t_1]]$

The structure in (24) shows the phrase *TIPOTA* "n-thing" raising from within the verb phrase to the clause edge, where it marks its scope over the negative marker *dhen*. Crucially, the phrase *TIPOTA* is not itself semantically negative. Since the marker *dhen* contributes the only semantic negation in the structure, the single negation NC reading is derived.

From the perspective of this paper, the importance of Giannakidou's (1999, 2000) theory is the clear link established between NC and NPI constructions. However, along with theories such as Zeijlstra (2004, et seq.), it is difficult to extend to grammatical systems that generate both NC and DN (e.g., Déprez, 2011; Puskás, 2012; Déprez et al., 2015, among others), including English (e.g., Blanchette and Lukyanenko, 2019). For example, if English negative phrases are NPI-like, then they should not be able to occur in DN constructions. A further prediction is that languages with NC should not have negative phrases appearing with no accompanying clause-bound negative marker, but as Tubau (2016) shows, such sentences coexist in vernacular Englishes alongside NC [see (9) and (10) above], and as we demonstrate below in our experimental results, the same appears to be the case for SE[16].

Strong vs. Weak NPIs

Zwarts (1998) observes within-language diversity in NPI licensing patterns, which serves as the basis for the two syntactic conditions we employ in our experiment. Consider the following examples:

(25) Maria didn't eat anything for lunch today.
(26) Maria didn't eat a damn thing for lunch today.
(27) If Maria eats anything for lunch today, she'll be able to work through the afternoon.
(28) *If Maria eats a damn thing for lunch today, she'll be able to work through the afternoon.

Sentences (25) through (28) contain the NPIs *anything* and *a damn thing*. While *anything* is acceptable in both the negative context in (25) and the non-negative conditional context in (27), *a damn thing* is only acceptable in the negative context (26), and (28) is unacceptable. Zwarts characterizes this behavior in terms of NPI strength. NPIs such as *a damn thing* are strong, in that they require a strong licensing context such as negation. NPIs such as *anything* are weak, in that, while they are licensed under negation, they may also appear in semantically weaker contexts such as conditionals[17].

A Unified Syntactic Account of NPIs, NC, and DN

Postal (2005) diverges from previous accounts of NPI behavior in proposing that NPIs themselves introduce negation into the structure. Under his theory, there exist two possible underlying structures for NPIs, which Collins and Postal (2014) call "unary NEG" NPIs and "reversals," and which they propose map onto strong and weak NPIs respectively. The following are Postal's proposed structures for these two NPI types:

(29) Unary NEG NPI: $[_{DP}$ $[_D$ NEG SOME] X] ("strong" NPIs)

(30) Reversal: $[_{DP}$ $[_D$ NEG $[_D$ NEG SOME]] X] ("weak" NPIs)

Both structures are noun phrases (DPs) with a negation (NEG) directly modifying an abstract SOME.

Postal (2005) further proposes that NPIs with the forms *anything*, *anybody*, and the like, may have either a unary NEG or a reversal structure. When they occur with the unary NEG structure, the negation that is introduced within the NPI raises to a higher position in the syntax, as follows:

(31) Structure for 'Maria didn't drive any cars.'
Maria didNEG_1 drive [NEG_1 SOME cars]

Collins and Postal (2014) propose that the surface form for a structure such as (27) is derived when the lower copy of the

[13]See Herburger (2001, p. 295) for a similar conclusion.

[14]A detailed summary of Giannakidou's (1999) theory of NPI licensing as veridicality sensitivity is beyond the scope of this paper, but see Giannakidou (2000) and Liu (2019) for this, and see Liu (2019) for an experimental investigation of NPI licensing in conditionals which can be taken to support this theory.

[15]Capital letters denote emphasis, which according to Giannakidou (2000) is the property which distinguishes negative universal quantifiers from existentials.

[16]See also Agostini and Schwenter (2018) for corpus and experimental evidence of this phenomenon in Brazilian Portuguese.

[17]See Gajewski (2011) for a proposal in which NPI strength is explained by appeal to (non-)sensitivity to non-truth conditional aspects of meaning.

negation goes unpronounced and abstract SOME maps to surface form *any*. The structure in (31) thus derives the dependency between the NPI and the higher negation without appeal to semantic licensing.

Note now that the reversal structure in (30) has a second negation. Their proposal is that the outer negation cancels the force of the inner one, yielding a non-negative semantics. This model thus generates the correct truth conditions for sentences such as conditionals, in which NPIs are licensed. For example, in the sentence *If Maria drives any cars, she'll drive them fast*, the term *any* can be replaced by *some* (or removed entirely) with no change in truth conditions[18].

Blanchette (2015) uses data from Appalachian vernacular English to show how this system readily extends itself to NC. For an NC sentence like *Maria didn't drive no cars*, the structure is the same as in (31), except both copies of the negation are spelled out in the phonology, leaving abstract SOME unpronounced. For the DN interpretation (which also exists in Appalachian), the structure simply contains two distinct semantic negations, and there is no NEG raising to a higher position, hence no negative dependency is established:

(32) DN structure:
Maria didNEG_1 drive [NEG_2 SOME cars]
Meaning: It is not the case that Maria drove no cars. (= She drove at least one car.)

A further benefit of the Postal (2005) and Collins and Postal (2014) system is that it also captures data such as those observed in Tubau (2016), in which negative noun phrases appear variably in concord with a clause mate negative marker, and independently, with no negative clause mate, as in (9) and (10) above. The theory derives these by positing that a unary NEG noun phrase is present in the structure, but the negation remains in its base position and does not undergo raising.

The Current Study

In light of the English data examined here, a benefit of Postal (2005) and Collins and Postal's (2014) theory is that it allows for the generation of both NC and DN structures alongside NPI constructions, within the same grammatical system, while previous syntactic and semantic accounts these phenomena do not yet have a clear answer for how all of this might work together. While the current study is not designed to test a particular theory, it does explore the degree to which the same population of speakers treats sentences with overtly negative noun phrases and NPI constructions as parallel, and therefore, the extent to which it is desirable to model them in the same way. We sought to find experimental evidence to support the idea that speakers calculate parallel truth conditions for NC and NPI constructions with negative marker, a "strong" licensing context (and both underlyingly unary NEG structures according to Collins and Postal, 2014 and Blanchette, 2015), and concurrently, whether these same speakers understand that the semantic contributions of the NPI and negative noun phrase yield opposite truth conditions in conditionals, a "weak" and non-negative NPI licensing context (and a context for reversals under Collins and Postal, 2014). As we will show below, the experiment design works because of the nature of the NPI itself. Specifically, when in the scope of a negation, the NPI shares a meaning with the overtly negative noun phrase in NC, but when in the "weak" reversal context of a conditional it takes on the opposite meaning, which is logically non-negative.

[18]Collins and Postal (2014, Chapter 8) derive the surface forms of reversal NEG structures by proposing a system of "NEG deletion" which involves a relationship between the inner and the outer NEG, and the outer NEG and a "NEG deleter" that structurally precedes it. The process of NEG deletion removes both negations from the phonological output.

MATERIALS AND METHODS

Research Questions

Our experiments were designed to explore similarities and differences between overtly negative noun phrases and NPIs in direct object position under negation, a context for "strong" NPI licensing or unary NEG NPI structures, and under conditionals, a context for "weak" NPI licensing or reversal structures (Zwarts, 1998; Postal, 2005; Collins and Postal, 2014). We asked the following questions:

(i) Do English speakers access parallel meanings for NPI and NC constructions under negation (i.e., contexts for unary NEG structures), despite asymmetries in the acceptability of these constructions?
(ii) Do these same English speakers readily distinguish between the meanings of NPIs and overtly negative noun phrases in "weak" (reversal) licensing contexts, which do not parallel NC?

Participants

Thirty participants (10 women, 20 men) were recruited through Amazon Mechanical Turk (AMT) for the main experiment, and a further 15 (5 women, 10 men) were recruited for the follow-up. To participate, speakers had to confirm that they were at least 18 and spoke American English natively. Completing the online survey took approximately 30 min, and participants were paid $6 for their time.

All participants had spent most or all of their lives in the US. Their answers to free response questions about cities and regions where they had lived indicated that 17 had spent the majority of their childhoods in the south (including 4 in Florida and 3 in Texas), 9 in the Midwest, 7 in the midatlantic, 6 on the west coast, 1 each in the northeast, great plains and southwest, and that 3 had spent similar amounts of time in two or more regions. Four participants reported familiarity with a language other than English, two heritage language speakers (Chinese, Spanish), and two foreign language learners (Spanish, German).

Participants were between 24 and 72 years old (main study mean = 38.5 years, follow-up study mean = 40.2 years), and the majority had completed either high school (n = 9), or a 2-year (n = 10) or 4-year college degree (n = 16). Of the remaining participants, 5 had completed a graduate degree, 4 had begun a bachelor's degree, and 1 had begun a graduate degree.

An additional 5 participants, 4 from the main experiment and 1 from the follow-up, completed the task and were paid, but were excluded from the final dataset for failing to achieve 80% accuracy on the catch trials [described below, see (31g)]. These participants gave ratings of 5 or higher (i.e., felicitous) to 4 or more of the 16 fillers that were designed to have infelicitous continuations, or gave ratings of 4 or lower (i.e., unacceptable) to 4 or more of the first clauses of these fillers, despite the fact that these first clauses were unremarkable English sentences. This indicated either that they were not reading carefully, or that they had misunderstood the task.

In a post-survey language questionnaire, participants were asked how likely they and their family and friends were to use NC and NPI constructions to communicate a negative meaning, on a scale from 1 (never) to 7 (always). Ratings were low for use of NC (participants' median = 1, mean = 1.84; family and friends median = 2, mean = 2.6), and high for use of NPI constructions (participants' median = 6, mean = 5.9; family and friends median = 6, mean = 6.0). Given the heavy social stigma associated with NC, we interpret these responses with caution, but they suggest that the group of speakers who participated in our experiments can be characterized as primarily non-NC users.

Materials and Design

We designed two experiments to explore our research questions. The main experiment compared participants' ratings of three noun phrase types in negative contexts (a context for "strong" NPIs, or unary NEG structures) and in conditionals (a context for "weak" NPIs or reversals). The follow-up experiment further explored the acceptability of negative noun phrases in conditionals, and participants' interpretation of NPIs in these non-negative contexts. Both experiments included 48 critical sentences and 112 fillers. All sentences contained two clauses, the second of which described a consequence of or context for the first. See **Supplementary Appendix A** for a full list of items and fillers.

For critical sentences in the main survey, the first clause was either conditional or negative, and the direct object was a DP of one of three types: bare plural (*people, things*), NPI (*anybody, anything*), or negative noun phrase (*nobody, nothing*). Conditional clauses were followed by consequences consistent with a no-negation meaning, and negative clauses were followed by consequences consistent with a single negation meaning. DP type and sentence type were fully crossed within participants such that an individual participant saw 8 items in each of the six conditions, and never saw more than one form of a given item. Half of the items in each condition had animate direct objects (i.e., *people, anybody, nobody*), and half had inanimate objects (i.e., *things, anything, nothing*). Across participants, each item appeared equally in each condition, in a Latin Square design. Example sentences are shown in (33).

(33) a. If my older sister leaves things in her locker, then her backpack is gonna be a bit lighter during her walk home. *conditional-bare plural*
b. If my older sister leaves anything in her locker, then her backpack is gonna be a bit lighter during her walk home. *conditional-NPI*
c. If my older sister leaves nothing in her locker, #then her backpack is gonna be a bit lighter during her walk home[19]. *conditional-negative noun phrase*
d. My older sister didn't leave things in her locker, so her backpack is gonna be super heavy during her walk home. *negative-bare plural*
e. My older sister didn't leave anything in her locker, so her backpack is gonna be super heavy during her walk home. *negative-NPI*
f. My older sister didn't leave nothing in her locker, so her backpack is gonna be super heavy during her walk home. *negative-negative noun phrase*

We were particularly interested in the comparison between NPIs and negative noun phrases in conditional and negative contexts, since this would show us whether speakers calculate parallel truth conditions for NC and NPI constructions in "strict," unary NEG contexts, and whether these same speakers also calculate opposite truth conditions when these noun phrase types appear the "weak" reversal context of conditionals. Constructions with a bare plural, which have the same truth conditional meaning as the NPIs in these sentences but no linguistic dependency, were employed as a control.

Critical sentences in the follow-up survey were derived from those in the main survey, by pairing the conditional first clauses with the single-negation continuations, as shown in (34). This was intended to render the negative noun phrases fully felicitous in the conditional sentences, and the NPIs and bare plurals infelicitous. Because there were only three conditions, participants saw twice as many sentences per condition as in the main study, and only half as many participants were needed to obtain the same number of observations per condition.

(34) a. If my older sister leaves things in her locker, #then her backpack is gonna be super heavy during her walk home. *conditional-bare plural*
b. If my older sister leaves anything in her locker, #then her backpack is gonna be super heavy during her walk home. *conditional-NPI*
c. If my older sister leaves nothing in her locker, then her backpack is gonna be super heavy during her walk home. *conditional-negative noun phrase*

Fillers were identical for the two surveys and were designed with the same two-clause structure as critical sentences. They included a variety of features intended to blend with the critical items, including several different subordinating conjunctions, universally quantified direct objects, and a single negated auxiliary without quantified or bare plural direct objects, as shown in (35). Of the 112 fillers, 96 were designed to have felicitous continuations (35a-f), and the remaining 16 were designed to be fully acceptable but have infelicitous

[19]The # symbol marks infelicity. We include it here for expository purposes. It was not included in the actual experiment.

continuations (35g). This created a similar proportion of infelicitous continuations in the filler items as we predicted there would be in the critical items (1/7 and 1/6 respectively). The 16 "mismatch" fillers served as catch trials and allowed us to exclude participants who had misunderstood the task or were not attending it carefully ($n = 5$, see section Participants).

(35) a. The playful kids left blocks all over the floor, so their parents are gonna make them clean up before dinner. *so*
b. If the strong wind blows the snow into the road, then drivers are gonna need to be careful coming through. *if-then*
c. The pro athlete is skipping her normal morning shower, because she's gonna go on a long run right after breakfast. *because*
d. The shy kitten hides behind the sofa whenever guests come over, but she's probably gonna come out later when it's dinner time. *but*
e. The taxi driver told everybody how dangerous the area was, so they're all gonna try to avoid it when they go out at night. *everybody/everything*
f. The teacher didn't open the window during the exam, so the students are all gonna be falling asleep in the heat. *single negation*
g. The highschooler received a perfect score on a really hard exam, #so his parents are gonna be really angry with him when he gets home. *mismatch/catch trial*

Procedure

Upon selecting the survey, AMT workers were directed to a Qualtrics survey link. They first read and acknowledged an informed consent statement, then proceeded to the survey[20]. For each item in the survey, participants were asked to first judge the naturalness (acceptability) of the first clause, and then judge the plausibility (felicity) of the second clause[21]. The targeted clause was bolded during the relevant judgment, but the entire sentence was visible throughout the trial. Both judgments were on a 7-point Likert scale, with endpoints labeled "completely natural" (7) and "completely unnatural" (1) for the acceptability rating, and "consequence makes total sense" (7) and "consequence makes zero sense" (1) for the felicity rating.

The survey was preceded by four practice trials with feedback, to familiarize participants with the task. Of the 4 practice trials, two had low acceptability first clauses (glaring word order errors), and two had high acceptability first clauses. This was crossed with plausibility of the consequence, to demonstrate the independence of the two judgments.

The body of the survey included the 112 fillers and 48 critical items presented in a fully random order and was followed by a short debriefing and language history questionnaire. Upon completion of the survey, participants were given a code to enter into the AMT interface in order to get their payment.

Analyses

Both acceptability and felicity ratings were on a 7-point Likert scale, and were therefore analyzed using ordinal rather than linear regression techniques (Liddell and Kruschke, 2018). All models were cumulative link mixed models, fit using the *clmm()* function of the *ordinal* package (version 2019.4-25; Christensen, 2019) in R (version 3.6.0; R Core Team,, 2019) and a probit link function.

This analysis technique differs in several ways from other common approaches to analyzing Likert data. Most importantly, in contrast to linear modeling techniques, ordinal modeling does not make the assumption that participants treat the ratings as equally spaced. That is, ordinal modeling allows for the possibility that participants will, for instance, be particularly hesitant to give the minimum rating, effectively making the distance between 1 and 2 larger than the distance between 2 and 3. Second, raw ratings are entered into the model, rather than z-scored ratings. Z-scoring serves two purposes when analyzing ratings using linear models: to make the measure more continuous and therefore more appropriate for a non-ordinal analysis, and to factor out between-participant variation. The use of ordinal analysis obviates the need for continuity, and mixed model approaches, whether linear or ordinal, take between-participant variation into account using random effects.

When interpreting model output, note that estimates are threshold changes in terms of shared standard deviation. Thus, while they are not readily interpretable as predicted change in score or probability (as might be the case in a well-coded linear model), they are readily comparable to each other within a model: an estimate of 3 indicates that a factor has twice as large an influence on the thresholds as a factor with an estimate of 1.5.

For other linguistic studies applying cumulative link mixed models to Likert scale judgment data, see Clifton et al. (2019), Fekete et al. (2018), and Scontras et al. (2017).

RESULTS

To explore the relationship between participants' acceptance of English NC and their ability to interpret it as truth conditionally equivalent to negative NPI constructions, we compared participants' acceptability ratings of three types of direct object (overtly negative noun phrases, NPIs, bare plurals) in negative and conditional sentences. Each initial clause was followed by a second clause that, for the negative sentences, was compatible with a single negation reading, and for the conditional contexts was compatible with a no-negation reading. We predicted that participants would rate all first clauses as relatively acceptable except for the negative noun phrase in a *negative* sentence, i.e., the stigmatized NC construction. We furthermore predicted that they would rate the consequence as highly felicitous for all second clauses except the negative noun phrase in a *conditional* sentence, which is incompatible with the meaning expressed in the consequence.

[20]This survey was conducted under the supervision of the Penn State IRB, which deemed it to be minimal risk and therefore exempt from requirements for written documentation of informed consent. Participants indicated their understanding of the consent document and willingness to participate by simply continuing with the survey.

[21]Previous studies have similarly elicited two separate responses of a different nature for a single item. See for example, Blanchette et al. (2018) and Li et al. (2019).

In the follow-up survey, we paired the single-negation compatible consequences with the conditional first clauses [see (30)] in order to confirm that negation is not uniformly less acceptable in conditionals, and that participants treat NPIs and negative noun phrases as opposites in non-negative conditional statements, in contrast to the negative contexts in the main study where we predicted they would be treated as syntactic variants.

Crucially, we predicted a disconnect between participants' acceptability ratings for NC sentences, and their felicity ratings for the single negation continuations in the main study, which would be instantiated as low acceptability but high felicity ratings for negative sentences with negative noun phrases. High felicity ratings would indicate that participants readily achieved the intended reading of the NC construction, and would suggest that low acceptability ratings are likely more the result of social pressure than the speaker's inability to generate the structure.

The Main Experiment

Figure 1 shows jittered raw ratings of sentence acceptability (left panel) and consequence felicity (right panel), along with boxplots to help summarize the distribution. The most striking pattern is the predicted reversal of the sentence type effect on negative noun phrases across the two panels. Negative noun phrases were rated as relatively unacceptable (median = 3) in negative contexts (the stigmatized NC construction), but their continuations, consistent with the single negation NC reading, were rated as highly felicitous (median = 7). In contrast, negative noun phrases in conditional sentences were rated as acceptable (median = 6), but their no-negation continuations were rated as infelicitous (median = 1). That is, participants appear to have rated stigmatized NC constructions as unacceptable, but readily generated the single negation interpretation necessary to make the consequence felicitous[22]. This supports the hypothesis that these constructions are part of the participants' grammars, but that their acceptability rating is heavily influenced by social pressure, and therefore serves as a poor diagnostic for grammaticality.

Other patterns visible in the graph include very high acceptability ratings for both bare plurals and NPIs in both conditional and negative sentences (all medians = 7), with the most consistently high acceptability ratings for bare plurals in conditional sentences, and generally high felicity ratings for consequences following bare plurals and NPIs (median = 6 for conditional-NPI, 7 elsewhere). Also note that there is more spread in the generally low ratings for negative NPs in the negative sentences (median = 3) than one might expect for something truly unacceptable. Compare, for instance, the consistent, very low felicity ratings (median = 1) for the truly infelicitous continuations, following conditional sentences with negative noun phrases. We return briefly to this variability in the discussion.

To explore these patterns statistically, we fit separate cumulative link mixed models for acceptability ratings and felicity ratings (see section Analyses). For both models, predictor variables were the two-level factor sentence type (conditional, contrast code -0.5 vs. negative, contrast code 0.5), and the three-level factor NP type, coded as two Helmert contrasts, the first comparing negative noun phrases to NPIs and bare plurals together ("negative-other," negative noun phrases, 0.67 vs. NPIs

[22]We are currently collecting data in a parallel eye-tracking study. Tracking participants' eye-movements as they read these same stimuli will allow us to more directly investigate just how readily this interpretation is generated.

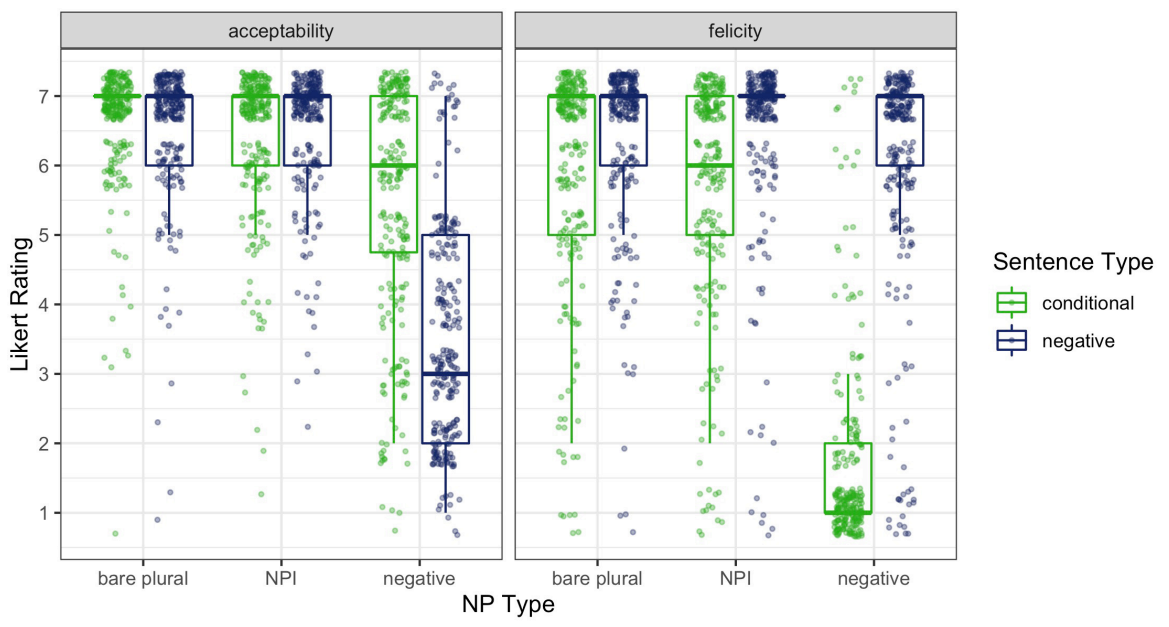


FIGURE 1 | Raw acceptability and felicity judgments for the main survey on a 7-point Likert scale, with box plots showing overall quartiles and median. Values for conditional sentences are shown in light green, and for negative sentences in dark blue.

and bare plurals, both -0.33), and the second comparing NPIs to bare plurals ("NPI-bare," NPIs, -0.5 vs. bare plurals, 0.5, negative NPs, 0), as well as the interactions of sentence type and the NP type contrasts. The model included random intercepts for item and participant and the random slopes of sentence type by item and of NP type, sentence type and their interaction by participant.

Acceptability

Model results are shown in **Table 1**. All main effects were reliable, as was the key interaction of sentence type and the negative-other NP type contrast [all LR(1) > 4, all $p < 0.05$]. This reliable interaction is consistent with our expectation that negative noun phrases in negative contexts would be treated as particularly unacceptable.

Planned comparisons further explored the key interaction and supported this conclusion. Three models, identical to the main model except for their contrast codes, were conducted to examine the simple main effects of both NP type contrasts in conditional and in negative sentences, and to examine the simple main effect of sentence type on the acceptability of negative noun phrases. These models revealed that negative noun phrases were less acceptable than NPIs and bare plurals in negative sentences [$b = -3.27$, se = 0.28, LR(1) = 57.65, $p < 0.00001$] and somewhat less acceptable (note the much smaller estimate) in conditional sentences [$b = -1.52$, se = 0.19, LR(1) = 38.5, $p < 0.00001$]. The interaction in the main model indicates that this difference was reliably larger for negative sentences than conditionals, and a follow-up comparison confirms that negative noun phrases in negative sentences (i.e., NC constructions) were reliably less acceptable than in conditional sentences [$b = -1.73$, se = 0.21, LR(1) = 38.82, $p < 0.00001$]. Intriguingly, NPIs were also very slightly but reliably less acceptable than bare plurals in conditional sentences [$b = 0.59$, se = 0.21, LR(1) = 8.36, $p = 0.004$], but not in negative sentences [$b = 0.06$, se = 0.24, LR(1) = 0.07, $p = 0.79$], perhaps reflecting the additional processing load incurred by the interaction between the NPI and the conditional. We discuss this further below.

Felicity

For felicity ratings, the continuations of the conditional sentences with negative noun phrases were predicted to be infelicitous, which should result in a reliable interaction between sentence type and the negative-other NP type contrast. This prediction was supported by the model results, shown in **Table 2**. All main effects and the interaction of sentence type with the negative-other NP type contrast were statistically reliable [all LR(1) > 7, all $p < 0.01$]. The interaction of the negative-other contrast and sentence type supports our prediction that negative noun phrases in conditional contexts would be treated as particularly infelicitous.

TABLE 1 | Model results for acceptability ratings in the main survey.

Effect	Estimate	se	z	LR (1)	p
NP type					
negative-other	−2.40	0.20	−11.78	57.85	<0.00001
NPI-bare	0.33	0.16	2.03	4.05	0.044
Sentence type	−0.57	0.13	−4.21	14.27	0.0002
Sentence type × NP type					
negative-other	−1.75	0.25	−7.14	33.26	<0.00001
NPI-bare	−0.52	0.31	−1.68	2.65	0.10

All p-values were obtained using likelihood ratio tests.

TABLE 2 | Model results for felicity ratings in the main survey.

Effect	Estimate	se	z	LR (1)	p
NP type					
negative-other	−1.64	0.13	−12.69	57.46	<0.00001
NPI-bare	−0.05	0.10	−0.49	0.25	0.62
Sentence type	1.51	0.15	9.91	51.02	<0.00001
Sentence type × NP type					
negative-other	2.26	0.29	7.85	43.09	<0.00001
NPI-bare	−0.51	0.19	−2.65	7.37	0.007

All p-values were obtained using likelihood ratio tests.

This primary model was again followed by further analyses to explore the interactions in the data. These revealed a reliable effect of the negative-other NP type contrast in both conditional [$b = -2.77$, se = 0.20, LR(1) = 64.57, $p < 0.00001$] and negative sentences [$b = -0.51$, se = 0.19, LR(1) = 5.84, $p = 0.02$]. The former supports the predicted interaction, and the latter indicates that while negative noun phrases were felicitous under negation (with median acceptability of 6), they were reliably less felicitous than NPIs and bare plurals. Further supporting the predicted interaction, we found that continuations of conditional sentences with negative noun phrases were reliably less felicitous than continuations of negative sentences with negative noun phrases [$b = 3.01$, se = 0.30, LR(1) = 46.58, $p < 0.00001$]. This indicates that participants reliably distinguished between NC sentences which made the continuation felicitous, and conditional *if*-clauses which did not.

Again, intriguingly and consistent with the overall interaction between sentence type and the NPI-bare contrast, there were differences between NPIs and bare plurals. There was a reliable effect of the NPI-bare contrast in negative sentences [$b = -0.30$, se = 0.16, LR(1) = 4.13, $p = 0.04$] and a marginal one in the opposite direction in conditional sentences [$b = 0.20$, se = 0.11, LR(1) = 2.90, $p = 0.09$]. That is, continuations were reliably rated as more felicitous for bare plurals in conditionals and for NPIs in negative sentences, perhaps reflecting a greater ease of processing NPIs in negative ("strict," or unary NEG) contexts than in non-negative conditional ("non-strict," or reversal) contexts.

The Follow-Up Experiment

One possible explanation for the pattern of felicity ratings in the main study is that the consequences of conditional sentences with negative direct objects were rated as infelicitous at least partly because negation is difficult to process, and this was exacerbated by the presence of the conditional. The follow-up survey was designed to confirm first that sentences with negative noun phrases were not inherently less felicitous under conditionals, and further, to confirm that speakers understand when negative noun phrases are equivalent in meaning to NPIs and when they are not. **Figure 2** shows participants' raw acceptability and felicity ratings

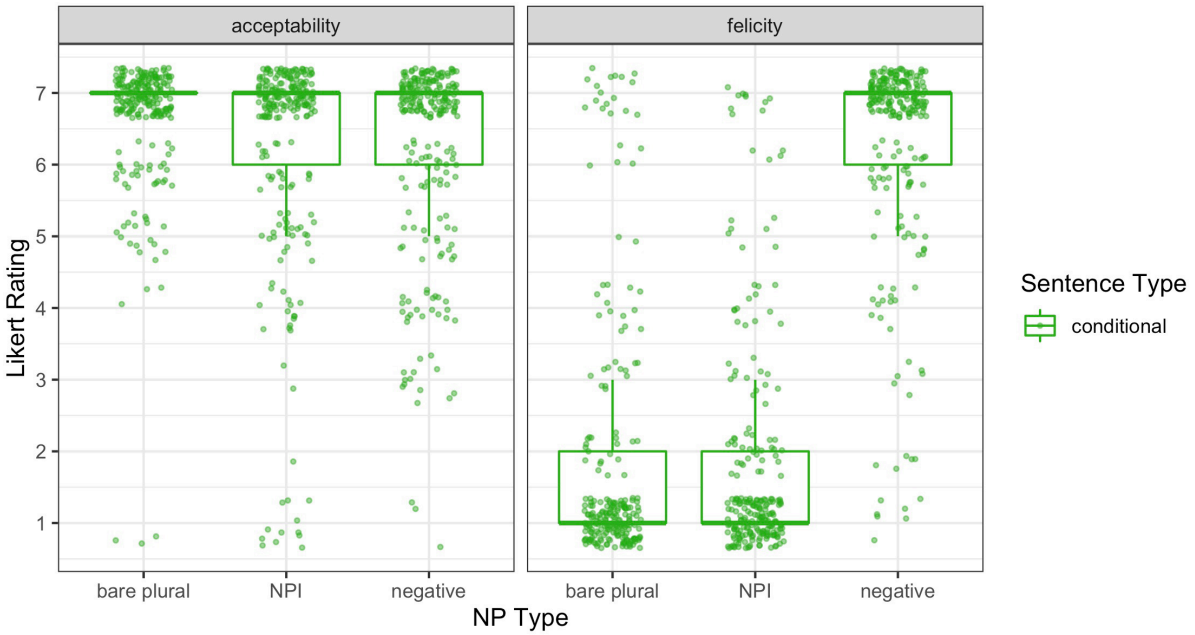


FIGURE 2 | Raw acceptability **(left panel)** and felicity ratings **(right panel)** for the follow-up survey on a 7-point Likert scale, with box plots showing overall quartiles and median.

for the sentences in the follow-up survey. Acceptability was very high across all NP types (all medians = 7), and felicity of the single negation consequence was rated as very low for the bare plural and NPI sentences (medians = 1), and very high for the negative NP sentences (median = 7).

We again fit cumulative link mixed models of acceptability ratings and of felicity ratings, this time with a single fixed effect predictor, the three-level Helmert-coded factor NP type. The models had random intercepts for participants and items, and random slopes for NP type by participant. For acceptability ratings, the model revealed only a marginal main effect of the NPI-bare NP type contrast (model results are shown in **Table 3**), reflecting the slightly higher ratings for bare plurals relative to NPIs and replicating the pattern found in follow-up analyses in the main study. There was no reliable decrease in acceptability for negative noun phrases as compared to the other NP types.

For felicity ratings, there was no reliable difference between bare plurals and NPIs, but there was a reliable difference between negative NPs and the other NP types (model results are shown in **Table 4**). This again confirms that in conditional sentences, negative (or unary NEG) noun phrases contribute a negative meaning, rendering the single-negation compatible continuation felicitous, while NPIs (here, reversals) contribute a meaning truth conditionally equivalent to non-negative bare plurals.

TABLE 3 | Model results for acceptability ratings in the follow-up survey.

Effect	Estimate	se	*z*	LR (1)	*p*
NP type					
negative-other	−0.31	0.27	2.08	1.21	0.27
NPI-bare	0.51	0.25	−1.15	3.53	0.06

All p-values were obtained using likelihood ratio tests.

TABLE 4 | Model results for felicity ratings in the follow-up survey.

Effect	Estimate	se	*z*	LR (1)	*P*
NP type					
negative-other	2.67	0.26	10.43	33.07	<0.00001
NPI-bare	−0.01	0.13	−0.09	0.01	0.93

All p-values were obtained using likelihood ratio tests.

DISCUSSION

Our main experiment involved two comparisons, one which compared negative noun phrases, NPIs, and bare plural controls under negation, and another which compared these same elements under non-negative conditionals. We asked whether speakers would calculate parallel truth conditions for NPIs and overtly negative noun phrases under negation (a context for unary NEG structures), and whether these same speakers would calculate opposite truth conditions for these words in non-negative conditionals (a context for reversal structures). We first discuss the comparison which involved negative dependencies.

Negative Contexts

Comparison of the three argument types in syntactically negative contexts revealed an asymmetry which can inform our understanding of the relationship between NC and NPI constructions: Participants' acceptability ratings of socially stigmatized NC constructions were low, and their felicity ratings

of consequences which correspond to the NC interpretation for these same constructions were high. Furthermore, while NC and NPI constructions were rated on opposite sides of the scale in acceptability, with NC on the low side and NPIs on the high side (and similar to bare plural controls), the consequences of all construction types were given relatively high felicity ratings.

Regarding the asymmetry between NC acceptability and felicity, we note that this finding both supports and complements previous work which compared NC with DN, its truth conditional opposite (Blanchette, 2017; Blanchette et al., 2018; Blanchette and Lukyanenko, 2019). In these studies, preceding context was employed to elicit an NC or a DN reading for sentences a subset of which were parallel to the critical sentences presented here. Speakers were shown, through a variety of measures including naturalness ratings, forced choice judgments of meaning, acoustic production and perception, and online reading times, to reliably prefer the NC over the DN interpretation for these items.

In the current study, there were no DN interpretations elicited from speakers during the course of the experiment, and given that participants reliably judged the single negation consequence of NC constructions as felicitous (which would have been infelicitous on a DN reading), we can assume that, at least for the most part, participants did not generate DN meanings for the items with two syntactic negations. Other differences between this study and those previous studies include the fact that participants judged NPI and bare plural sentences as well as NC, and that their judgments were made on the basis of the NC interpretation's felicitousness as determined by a following consequence, as opposed to a preceding context. In the context of previous studies in which speakers reliably prefer NC over DN interpretations in a range of measures, the fact that a distinct design led to similar results thus further confirms the robustness of speakers' readiness to interpret NC constructions as singly negative, and provides complementary support for the hypothesis that speakers who do not accept NC nevertheless have grammatical knowledge of it.

Regarding the interactional aspect of the asymmetry in negative contexts, in which NC acceptability and felicity were at opposite sides of the scale, while NPI (and bare plural) acceptability and felicity were on the same side, this shows that participants readily accessed the same truth-conditional meaning for all three NP types under negation, despite reliable differences in their acceptability. It should be noted that there was in fact a small but statistically reliable difference between NPI and bare plural felicity in negative contexts on the one hand, and NC felicity on the other. We believe this difference is best explained as a carryover effect of the strong unacceptability of NC. This is particularly likely since, as explained in the methods section, participants still had the critical sentence in view when judging the consequence.

The interaction between NC and NPI constructions in negative contexts also illustrates a more general methodological point, namely, that examining acceptability in isolation from meaning can obscure speaker knowledge of a construction type (especially where that construction type is socially stigmatized). In this case, the social stigma associated with English NC appears to be a primary force shaping speakers' acceptability ratings. Yet despite the strength of this social stigma, participants drew a clear distinction between the acceptability of NC and its meaning in context. NC thus provides an example of a construction type for which binary or overall acceptability and interpretation are unrelated. We extend this to suggest that NC also provides an example of a construction type for which overall acceptability and *grammaticality* are unrelated, and participants are able to interpret NC structures because their grammars generate them. This means that, in the case of NC, the traditional direct link between acceptability and grammaticality fails. Below we discuss some theoretical implications of participants' readiness to assign the same meaning to NC and NPI constructions, despite their distinct acceptability status.

To conclude this subsection, we note that there was substantially more spread in the negative sentence-negative noun phrase (i.e., NC) acceptability ratings than what might be expected for something that is outright ungrammatical (e.g., sentences with glaring word order violations such as *Up the bike the woman the hill rode*). The median response for NC sentences is 3, and observing the individual data points in **Figure 1**, we see that there are also many 4s and 5s. Thus, while overall acceptability is significantly lower for these NC constructions than for their prescriptively correct variants, these middling acceptability ratings may hint at their hypothesized grammaticality. Another possibility is that, because a large proportion of the sentences within the experiment were acceptable, participants were more inclined to provide slightly higher ratings even for the least acceptable sentences. The latter interpretation maintains the conclusion that there is no relation between English NC acceptability and grammaticality, while the former suggests some potential overlap.

Conditional Contexts

Items where the NPIs, overtly negative noun phrases, and bare plurals appeared under conditional contexts displayed two clear additional asymmetries beyond the ones found in negative contexts. In the main experiment, the clearest asymmetry was again interactional in nature, between the NPIs and bare plurals on the one hand, and the overtly negative noun phrases on the other. These were all relatively acceptable, with mean scores well above the middle of the scale, but in the main study, the contexts were designed to make the NPIs and bare plurals felicitous and the overtly negative noun phrases infelicitous. Unsurprisingly, participants responded in reliable fashion to this experimental manipulation, rating consequences following *if* clauses with overtly negative noun phrases as extremely low, despite the relative acceptability of the *if* clause itself. Viewed alongside the behavior of NC and NPI constructions in negative contexts, what this asymmetry shows is that the same participants who understood that negative noun phrases and NPIs are truth conditionally equivalent in negative contexts (i.e., contexts for unary NEG structures) readily reversed the truth condition for NPIs in non-negative (i.e., reversal) contexts.

The follow-up experiment was designed to inform the results of the main experiment, and to provide a more complete picture of speakers' understanding of where contexts for NPIs and overtly

negative noun phrases do and do not overlap. Reversing the truth conditions for the consequence from the main experiment, we expected that the non-negative NPI (a reversal structure), and not the (unary NEG) negative noun phrase, would be infelicitous. Participants again behaved as predicted, rating consequences of NPI and bare plural *if* clauses at the very low end of the scale, and consequences of overtly negative noun phrases at the high end. This allows us to point to the consequence as the source of infelicity for the negative noun phrase in conditionals in the main experiment. Additionally, it confirms that speakers understand when the meaning of an NPI is equivalent to an overtly negative noun phrase which participates in concord, and when it is not.

Before turning to theoretical implications, we note an additional asymmetry that our experiment was not explicitly designed to reveal: Though acceptable overall, overtly negative noun phrases were slightly less acceptable than NPIs and bare plurals in the main experiment conditional contexts. One potential explanation for this is that negation makes things more difficult to process (e.g., Ferguson et al., 2008), thus degrading acceptability, and further, that participants prefer a more focalized information status for negative objects with no preceding negative marker (e.g., Childs, 2017; Palacios Martínez, 2017). Note, however, that when the consequence for *if* clauses with an overtly negative noun phrase object was made felicitous, as in the follow-up experiment, the median acceptability of *if* clauses with NPIs and those with overtly negative noun phrases was nearly identical. It is therefore more likely that the infelicity of the consequence carried over here in the reverse direction, degrading the acceptability of the *if* clause where the object was overtly negative. This conclusion is supported by the fact that NPI acceptability in *if* clauses was on par with negative noun phrase acceptability in the follow-up experiment. Interestingly, this degradation effect did not apply to the bare plurals in the follow-up experiment. This suggests a potentially interesting conclusion that the source of this degradation is the negative dependency itself, suggesting that the cost of processing this dependency impacts acceptability ratings. Alternatively, it might be the case that the presence of heavily stigmatized NC in the main experiment served to degrade participants' acceptability judgments of all sentences with negative noun phrases. We leave this matter for future research.

Theoretical Implications

One explanation for the fact that participants gave similar felicity judgments for the NC and NPI constructions in negative or "strict" contexts is that their grammars represent NC and NPI constructions as syntactic variants with the same underlying form. This explanation finds its theoretical basis in Postal (2005) and Collins and Postal's (2014) analysis of NPI constructions, and Blanchette's (2015) extension of this proposal to English NC, described above. Under this theory, the grammar of the negative NPI and the NC constructions in this experiment involve the raising of a negation from the object noun phrase to a higher clausal position, generating a syntactic dependency between the negative marker and the object. The only difference between the two constructions is at the level of phonological spell out: In the one that surfaces as an NPI construction, the negation is unpronounced (and an abstract SOME spells out as *any*), whereas the NC construction involves spell out of both negations (and a silent abstract SOME).

The process governing the spell out of the lower negation in unary NEG structures may be grammatical in nature, where SE grammars have a constraint that prohibits them from pronouncing the lower negation which is absent from vernacular varieties, or it may be a purely socially governed phenomenon which over time has been conventionalized in SE, with the effect of masking a direct underlying grammatical connection between these two construction types. Whether the differences between these two surface forms are derived by grammatical or social pressures, a plausible explanation for our results is that speakers generated the same negative dependency in both the NC and the NPI constructions in negative contexts, and this was reflected in their felicity judgments. Concurrently, their clear intuitions about the opposite meanings of negative noun phrases and NPIs in conditionals, a "weak" licensing context, supports the hypothesis that they also have two distinct underlying representations for NPIs, a unary NEG structure and a semantically non-negative reversal, and they select the item analogous to the reversal structure for these conditional contexts.

We can also view our results in light of Tubau's (2016) theory of English NC. The extension would be similar to that of Collins and Postal (2014) in the sense that it would also assume speakers have two lexical entries for the same word, except that, instead of having two entries for *any*-NPIs, there would be two distinct entries for overtly negative noun phrases, one of which appears in NC constructions, and one of which appears in conditionals. We would then need to extend the theory further to account for the behavior of NPIs, and specifically, to explain not just the dependencies involved in these, but also, why they overlap in meaning with NC constructions in negative contexts, but contribute a meaning that reverses the truth conditions for the negative noun phrase in conditionals.

With regard to purely semantic theories of NPI licensing, in addition to finding experimental evidence for a parallel to the calculation of downward entailing inferences (Ladusaw, 1979), or (non-)veridicality (Giannakidou, 1998, 1999) in processing, we would now also need to explore whether the dependency established in NC, coupled with the now well-established observation that NC and DN may coexist in a single system, can also be explained by these theories. We set these questions, and the design of more targeted experiments which can tease apart these theories of grammar, aside for future work.

CONCLUSION

The experiments we reported here revealed asymmetries in the acceptability and felicity of NC and NPI constructions. We have provided evidence that speakers understand when the truth conditions for NC and NPI constructions overlap, and when they do not. The results have both methodological and theoretical implications. On the methodological side, they

demonstrate a clear case where there is no straightforward causal link between acceptability and grammaticality, and concurrently how judgments of meaning can inform theories of grammar in cases where acceptability judgments fail. On the theoretical side, they show how the set of facts that grammatical theories should be capable of modeling within a single system includes NC and NPI constructions, and in the context of previous studies, also DN. We further discussed how the system in Postal (2005) and Collins and Postal (2014), and its extension in Blanchette (2015), provides one such theory, while other existing theories do not yet explicitly capture the full range of facts.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval were obtained for the study on human participants in accordance with the local legislation and institutional requirements, and the study was determined to be exempt from continuing review. Written informed consent to participate in this study was not required in accordance with the national legislation and institutional requirements.

AUTHOR CONTRIBUTIONS

Both authors made substantial contributions to the development of this work, including experimental design, data collection and analysis, drafting and revising the manuscript, and agreed to be accountable for all aspects of the work.

FUNDING

The funds that supported this research were provided by Penn State Eberly College of Science and the George Mason University Linguistics Program. The publication fee was provided by Penn State and the George Mason Libraries Open Access Publishing Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02486/full#supplementary-material

REFERENCES

Agostini, T., and Schwenter, S. (2018). Variable negative concord in Brazilian Portuguese: acceptability and frequency. *Issues Hispanic Lusophone Linguist.* 2015, 71–93.

Auwera, J., and Alsenoy, L. (2016). On the typology of negative concord. *Stud. Nat. Lang.* 40, 473–512. doi: 10.1075/sl.40.3.01van

Baker, C. L. (1970). "Problems of polarity in counterfactuals," in *Studies Presented to Robert B. Lees by his Students, PIL Monograph Series 1*, eds J. Sadock, and A. Vaneck, (Edmonton: Linguistic Research Inc.).

Barwise, J., and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguist. Philos.* 4, 159–219.

Bayer, J. (1990). "What Bavarian reveals about the syntactic structure of German," in *Glow Essays for Henk Van Riemskijk*, eds H. Hulst, J. Mascaro, and M. Nespor, (Berlin: De Gruyter), 13–24. doi: 10.1515/9783110867848.13

Biberauer, T., and Zeijlstra, H. (2012). Negative concord in Afrikaans: filling a typological gap. *J. Semant.* 29, 345–371. doi: 10.1093/jos/ffr010

Blanchette, F. (2013). Negative Concord in English. *Linguist. Variat.* 13, 1–47. doi: 10.1075/lv.13.1.01bla

Blanchette, F. (2015). *English Negative Concord, Negative Polarity, and Double Negation*. Ph.D. thesis, CUNY Graduate Center, New York, NY

Blanchette, F. (2016). "English negative concord and double negation: applying the framework of Collins and Postal 2014," in *NELS 46: Proceedings of the Forty-Sixth Annual Meeting of the North East Linguistic Society*, eds C. Hammerly, and B. Pickett, (Amherst, MA: Graduate Linguistics Student Association), 107–122.

Blanchette, F. (2017). Micro-syntactic variation in English negative concord. *Glossa* 2, 1–32.

Blanchette, F., and Lukyanenko, C. (2019). Unacceptable grammars? an eye-tracking study of English negative concord. *Lang. Cogn.* 11, 1–40. doi: 10.1017/langcog.2019.4

Blanchette, F., and Nadeu, M. (2018). Prosody and the meanings of English negative indefinites. *J. Pragmat.* 129, 123–139. doi: 10.1016/j.pragma.2018.03.020

Blanchette, F., Nadeu, M., Yeaton, J., and Déprez, V. (2018). English negative concord and double negation: the division of labor between syntax and pragmatics. *Proc. Annu. Meet. Linguist. Soc. Am.* 3, 1–15.

Chierchia, G. (2013). *Logic in Grammar*. Oxford: Oxford University Press.

Childs, C. (2017). Integrating syntactic theory and variationist analysis: the structure of negative indefinites in regional dialects of British English. *Glossa* 106, 1–31.

Christensen, R. H. B. (2019). *Ordinal - Regression Models for Ordinal Data. R Package Version 2019.4-25*. Available at: http://www.cran.r-project.org/package=ordinal/ (accessed April, 2019)

Clifton, C., Xiang, M., and Frazier, L. (2019). A note on the voice mismatch asymmetry in ellipsis. *J. Psychol. Res.* 48, 877–887. doi: 10.1007/s10936-019-09636-z

Collins, C., and Postal, P. M. (2014). *Classical NEG Raising: An Essay on the Syntax of Negation*. Massachusetts: MIT Press.

Collins, C., Postal, P., and Yeduvey, E. (2017). Negative polarity items in Ewe. *J. Linguist.* 54, 331–365.

Dayal, V. (1995). "Licensing any in non-negative/non-modal contexts," in *Proceedings of SALT V*, 72–93, eds M. Simons, and T. Galloway, (Ithaca, NY: Cornell University).

De Swart, H., and Sag, I. A. (2002). Negation and negative concord in Romance. *Linguist. Philos.* 25, 373–417.

Déprez, V. (2000). Parallel (a)symmetries and the internal structure of negative expressions. *Nat. Lang. Linguist. Theor.* 18, 253–342.

Déprez, V. (2011). "Atoms of negation: an outside-in micro-parametric approach to negative concord," in *The Evolution of Negation: Beyond the Jespersen Cycle*, eds R. Ingham, and P. Larrivée, (Berlin: Mouton de Gruyter), 221–272.

Déprez, V., Tubau, S., Cheylus, A., and Espinal, M. T. (2015). Double negation in a negative concord language: an experimental investigation. *Lingua* 163, 75–107. doi: 10.1016/j.lingua.2015.05.012

Déprez, V., and Yeaton, J. (2018). "French negative concord and discord: an experimental investigation of contextual and prosodic disambiguation," in *Proceedings of the Romance Languages and Linguistic Theory 14: Selected Papers from the 46th Linguistic Symposium on Romance Languages (LSRL)*, eds L.

Repetti, and F. Ordóñez, (Amsterdam: John Benjamins), 35–51 doi: 10.1075/rllt.14.03dep

Espinal, M. T., and Prieto, P. (2011). Intonational encoding of double negation in Catalan. *J. Pragmat.* 43, 2392–2410. doi: 10.1016/j.pragma.2011.03.002

Espinal, M. T., and Tubau, S. (2016). Interpreting argumental n-words as answers to negative questions. *Lingua* 177, 41–59. doi: 10.1016/j.lingua.2015.12.013

Espinal, M. T., Tubau, S., Borrás-Comes, J., and Prieto, P. (2016). "Double negation in Catalan and Spanish: interaction between syntax and prosody," in *Negation and Polarity: Experimental Perspectives*, eds P. Larrivée, and C. Lee, (Dordrecht: Springer), 145–176 doi: 10.1007/978-3-319-17464-8_7

Etxeberria, U., Tubau, S., Déprez, V., Borràs-Comes, J., and Espinal, M. T. (2018). Relating (Un)acceptability to interpretation. *Exp. Invest. Negat. Front. Psychol.* 8, 1–15. doi: 10.3389/fpsyg.2017.02370

Fekete, I., Schulz, P., and Ruigendijk, E. (2018). Exhaustivity in single bare wh-questions: a differential-analysis of exhaustivity. *Glossa* 3:96. doi: 10.5334/gjgl.549

Ferguson, H. J., Sanford, A. J., and Leuthold, H. (2008). Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Res.* 1236, 113–125. doi: 10.1016/j.brainres.2008.07.099

Gajewski, J. (2011). Licensing strong NPIs. *Nat. Lang. Semant.* 19, 109–148. doi: 10.1007/s11050-010-9067-1

Giannakidou, A. (1998). *Polarity Sensitivity as (non)Veridical Dependency*. Amsterdam: John Benjamins.

Giannakidou, A. (1999). Affective dependencies. *Linguist. Philos.* 22, 367–421.

Giannakidou, A. (2000). Negative concord? *Nat. Lang. Linguist. Theor.* 18, 457–523.

Giannakidou, A. (2002). "Licensing and sensitivity in polarity items: from downward entailment to nonveridicality," in *CLS 38: Papers from the 38th Annual Meeting of the Chicago Linguistic Society, Parasession on Polarity and Negation*, eds M. Andronis, A. Pycha, and K. Yoshimura, (Chicago, IL: University of Chicago).

Green, L. (2014). "Force, focus and negation in African American English," in *Micro-Syntactic Variation in North American English*, eds R. Zanuttini, and L. R. Horn, (New York, NY: Oxford University Press), 115–142. doi: 10.1093/acprof:oso/9780199367221.003.0004

Haegeman, L., and Zanuttini, R. (1996). "Negative concord in West Flemish," in *Parameters and Functional Heads: Essays in Comparative Syntax*, eds A. Belletti, and L. Rizzi (Oxford: OUP), 117–180.

Henry, A. (1995). *Belfast English and Standard English: Dialect Variation and Parameter Setting*. New York: Oxford University Press.

Herburger, E. (2001). The negative concord puzzle revisited. *Nat. Lang. Semant.* 9, 289–333.

Horn, L. (2010). "Multiple negation in English and other languages," in *The Expression of Cognitive Categories: Expression of Negation*, ed. L. Horn, (Berlin: Walter de Gruyter), 117–148.

Hudley, A. C., and Mallinson, C. (2010). *Understanding English Language Variation in U.S. Schools*. New York, NY: Teachers College Press.

Krifka, M. (1995). The semantics and pragmatics of polarity items. *Linguist. Anal.* 25, 209–257.

Labov, W. (1972). Negative attraction and negative concord in English grammar. *Language* 48, 773–818.

Ladusaw, W. A. (1979). *Polarity Sensitivity as Inherent Scope Relations*. New York, NY: University of Texas. Doctoral dissertation.

Lewis, S., and Phillips, C. (2015). Aligning grammatical theories and language processing models. *J. Psychol. Res.* 44, 27–46. doi: 10.1007/s10936-014-9329-z

Li, F., Borrás-Comes, J., and Espinal, M. T. (2019). Mismatches in the interpretation of fragments expressions in mandarin chinese. *J. Pragmat.* 152, 28–45. doi: 10.1016/j.pragma.2019.07.017

Liddell, T. M., and Kruschke, J. K. (2018). Analyzing ordinal data with metric models: what could possibly go wrong? *J. Exp. Soc. Psychol.* 79, 328–348. doi: 10.1016/j.jesp.2018.08.009

Linebarger, M. (1987). Negative polarity and grammatical representation. *Linguist. Philos.* 10, 325–387. doi: 10.1007/bf00584131

Liu, M. (2019). The elastic nonveridicality property of indicative conditionals. *Linguist. Vanguard* 5, doi: 10.1515/lingvan-2019-0007

Nevalainen, T. (2006). Negative concord as an English "vernacular universal": social history and linguistic typology. *J. Eng. Linguist.* 34, 257–278. doi: 10.1177/0075424206293144

Palacios Martínez, I. (2017). Negative concord in the language of British adults and teenagers English World-Wide. *J. Variet. Eng.* 38, 153–180. doi: 10.1075/eww.38.2.02pal

Parker, D., and Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition* 157, 321–339. doi: 10.1016/j.cognition.2016.08.016

Postal, P. (2005). "Suppose (if only for an hour) that negative polarity items are negation-containing phrases," in *Paper Presented at Read at Workshop on Polarity from Different Perspectives*, New York, NY: New York University

Prieto, P., Borràs-Comes, J., Tubau, S., and Espinal, M. T. (2013). Prosody and gesture constrain the interpretation of double negation. *Lingua* 131, 136–150. doi: 10.1016/j.lingua.2013.02.008

Progovac, L. (1994). *Positive and Negative Polarity: a Binding Approach*. Cambridge: Cambridge University Press.

Puskás, G. (2012). Licensing double negation in NC and non-NC languages. *Nat. Lang. Linguist. Theor.* 30, 611–649. doi: 10.1007/s11049-011-9163-z

R Core Team, (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Scontras, G., Polinsky, M., Tsai, C.-Y. E., and Mai, K. (2017). Cross-linguistic scope ambiguity: when two systems meet. *Glossa* 2, 1–28. doi: 10.5334/gjgl.198

Szabolsci, A., Bott, L., and McElree, B. (2008). The effect of negative polarity items on inference verification. *J. Semant.* 25, 411–450. doi: 10.1093/jos/ffn008

Thornton, R., Notley, A., Moscati, V., and Crain, S. (2016). Two negations for the price of one. *Glossa* 45, 1–30.

Tortora, C., Santorini, B., Blanchette, F., and Diertani, D. (2017). *The Audio-Aligned and Parsed Corpus of Appalachian English (*AAPCAppE). Available at: http://csivc.csi.cuny.edu/aapcappe/ (accessed August, 2019).

Tubau, S. (2016). Lexical variation and negative concord in traditional dialects of British English. *J. Comp. Germanic Linguist.* 19, 143–177. doi: 10.1007/s10828-016-9079-4

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Von Fintel, K. (1999). NPI licensing, Strawson entailment, and context dependency. *J. Semant.* 16, 97–148. doi: 10.1093/jos/16.2.97

Wallage, P. (2012). Negative inversion, negative concord and sentential negation in the history of English. *Eng. Lang. Linguist.* 16, 1–33.

Weldon, T. (1994). Variability in negation in African American English. *Lang. Variat. Change* 6, 359–397. doi: 10.1017/s0954394500001721

Wolfram, W., and Christian, D. (1976). *Appalachian Speech*. Arlington, VA: Center for Applied Linguistics.

Wolfram, W., and Fasold, R. W. (1974). *The Study of Social Dialects in American English*. Englewood Cliffs, NJ: Prentice-Hall, xii–239.

Zanuttini, R. (1997). *Negation and Clausal Structure: A Comparative Study of Romance Languages*. New York: Oxford University Press.

Zeijlstra, H. (2004). *Sentential Negation and Negative Concord*. Doctoral thesis, University of Amsterdam, Amsterdam

Zwarts, F. (1998). "Three types of polarity," in *Plurality and Quantification*, eds F. Hamm, and E. Hinrichs, (Dordrecht: Kluwer), 203–238.

Conflict of Interest: 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
published: 20 December 2019
doi: 10.3389/fpsyg.2019.02886

In Search of the Factors Behind Naive Sentence Judgments: A State Trace Analysis of Grammaticality and Acceptability Ratings

Steven Langsford[1]*, *Rachel G. Stephens*[2], *John C. Dunn*[3] and *Richard L. Lewis*[1]

[1] Department of Psychology, University of Michigan, Ann Arbor, MI, United States, [2] Department of Psychology, University of Adelaide, Adelaide, SA, Australia, [3] Psychological Science, University of Western Australia, Perth, WA, Australia

We present a state-trace analysis of sentence ratings elicited by asking participants to evaluate the overall acceptability of a sentence and those elicited by asking participants to focus on structural well-formedness only. Appealing to literature on "grammatical illusion" sentences, we anticipated that a simple instruction manipulation might prompt people to apply qualitatively different kinds of judgment in the two conditions. Although differences consistent with the subjective experience of grammatical illusion dissociations were observed, the state trace analysis of the rating data indicates that responses were still consistent with both judgment types accessing a single underlying factor. These results add to the existing comparisons between analytic and probabilistic modeling approaches to predicting rating judgments.

Keywords: acceptability, grammaticality, state trace analysis, rating task, language modeling

OPEN ACCESS

Edited by:
Susagna Tubau, Autonomous University of Barcelona, Spain

Reviewed by:
Dave Kush, Norwegian University of Science and Technology, Norway
Viviana Masia, Roma Tre University, Italy

***Correspondence:**
Steven Langsford
slangsfo@umich.edu

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

Received: *31 July 2019*
Accepted: *05 December 2019*
Published: *20 December 2019*

Citation:
Langsford S, Stephens RG, Dunn JC and Lewis RL (2019) In Search of the Factors Behind Naive Sentence Judgments: A State Trace Analysis of Grammaticality and Acceptability Ratings. Front. Psychol. 10:2886. doi: 10.3389/fpsyg.2019.02886

INTRODUCTION

Language communities have been shown to be consistent and reliable in their consensus reporting of how acceptable a sentence is (Sprouse et al., 2013; Mahowald et al., 2016). Quantifying, predicting, and contrasting ratings based on such judgments has for a long time been an important part of linguistics research (Schütze and Sprouse, 2014). But despite high agreement about *what is acceptable*, it is not at all obvious *what acceptability is*. Plausible candidates include the processing effort required (Braze, 2002; Hofmeister et al., 2013), the probability of the sentence under some appropriate language model (Chater and Manning, 2006), an expanded notion of probability including the naturalness/oddity given situational pragmatics (Masia, 2017; Domaneschi and Di Paola, 2018), or a combination of error signals that arise from different component stages of language processing (Sprouse, 2018).

Probably the most popular view is that acceptability is a combination of error signals from all these sources, which could of course include processing effort and word co-occurrence statistics as particularly salient signals (Sprouse, 2018). From this general perspective, a full understanding of acceptability ratings would entail describing the factor structure of linguistic acceptability and specifying how the different components interact.

Why Care About the Factor Structure of Ratings?

An understanding of the factor structure underlying sentence ratings may be helpful in interpreting conflicts between crowd-sourced acceptability judgments and sentence-status descriptions arrived at by analysis or other means. One such situation arises when crowd-sourced acceptability judgments conflict with descriptions of grammatical status arrived at by analysis or other means (Sprouse et al., 2013). How should these results be interpreted? Assuming that best practices have been followed to protect the reliability of the rating data (Myers, 2009; Ahler et al., 2019) one possible interpretation is that the analysis is in error. However, this is not the only interpretation. It is possible that the analysis and the acceptability judgment simply reflect distinct properties of the sentence, with acceptability responsive to a range of additional factors outside the scope of the analysis. Using structured interviews, Schütze (in press) finds that this is the case for at least some items identified by Sprouse et al. (2013) as examples of inconsistency between expert analysis and crowd-ratings. Schütze (in press) calls for qualitative data about the motivation for a rating to be collected alongside likert-style judgments, in order to identify the interpretation a rating was made under and any special features influencing the rating, such as an unknown word. It is possible that detailed instructions about the target property to be rated could reduce the variation in rating motivation. The study described below contrasts different instructions, giving an example of the size of such instruction-based effects.

Another arena in which the factor structure of judgment data is important is when it is used in the design and evaluation of language models. This usage could be direct, in a supervised learning system predicting acceptability ratings on hold-out items from a collection of rated sentences (Warstadt et al., 2019), or indirect, when the ability to predict sentence acceptability judgments is used to evaluate an unsupervised learning system trained on unannotated corpora (Lau et al., 2017). In either case, the composition of factors underlying ratings are important to the interpretation of the results. If sentence ratings are responsive to multiple properties of a sentence, for example both "surface probability" and "structural soundness," it is possible that evaluating models on their ability to predict ratings will lead to models that privilege one component at the expense of the other. A concrete example of this kind of feature-substitution appears in the computer vision literature, where convolutional neural nets have been found to weight texture more heavily than shape (Geirhos et al., 2018). This feature weighting is the exact opposite of the human pattern, but it arises naturally in this context because texture is highly predictive of object identity in the training data and involves short-range dependencies that are easier for these learning architectures to discover. To the extent that modern language modeling relies on similar learning architectures, it is similarly vulnerable to under-weighting or even omitting the "shape-like" properties of natural language if "texture-like" properties are available in rating judgments (Warstadt and Bowman, 2019).

One potential example of this scenario in linguistics is presented by Sprouse et al. (2018) in response to work by Lau et al. (2017) (see also Lappin and Lau, 2018). In brief, Sprouse et al. (2018) distinguishes between three different performance metrics in order to compare models presented by Lau et al. (2017) with existing theories of syntax as represented by submissions to *Linguistic Inquiry* and Adger's *Core Syntax* (Adger, 2003). One metric, the gradient metric, is a correlation between predicted rating and observed rating. Another, the categorical metric, is a discretized version of the gradient metric based only on the rank order of items. A third, the experimental-logic metric, counts successful predictions for the presence or absence of a difference in rating between two carefully controlled comparison items. The three performance measures are related: given a scheme for predicting rating scores for any sentence, the categorical and experimental-logic metrics are discretized versions of differences under the gradient metric. Despite this close relationship, Sprouse et al. (2018) report that high performance under the gradient metric is not necessarily associated with similarly high performance under the categorical and experimental-logic metrics. A striking feature of this work is the demonstration that categorical distinctions derived from the linguistic literature perform well on the two discrete metrics for which they are applicable but are not able to give predictions on the gradient metric, while a probabilistic model with attested high performance on the gradient metric shows a drop in performance when evaluated on the categorical and experimental logic metrics. One possible interpretation of this dissociation in performance might be that different linguistic properties are accessed by corpus-trained probabilistic models and expert analysis.

A second motivation for the study presented below is to explore contrasting explanations for what Sprouse et al. (2018) describe as a trade-off between performance on the gradient metric and the categorical metric. The suggestion that this performance trade-off reflects attention to different linguistic properties seems well-motivated on theoretical grounds, but in principle such dissociations can appear even if there is only one key well-formedness factor underlying both metrics (Loftus, 1978). In particular, we note that Bader and Häussler (2010) have explored a principled mapping between gradient and categorical judgments of grammatical status directly relevant to the distinction Sprouse et al. (2018) draws between the categorical and gradient evaluation metrics. The mapping scheme is an implementation of signal detection theory, and as such draws on a well-established tradition of such models in psychophysics (Green and Swets, 1966). This class of models contains a mechanism whereby responses can produce apparent dissociations even when both are based on the same latent factor (Stephens et al., 2018, 2019). Such an account would still be consistent with the performance contrasts demonstrated by Sprouse et al. (2018). Under a signal-detection account, a change in response thresholds, a change in noise levels, or both in concert could in principle produce differences like those observed between the discrete and gradient evaluation metrics even if both reflect a single underlying well-formedness factor. Proposing a single-factor account of differences between expert analyses

and probabilistic models may seem extreme given the extensive theoretical differences between these approaches. We raise the possibility to emphasize the way uncertainty about the factor structure of acceptability rating judgments leaves unclear what kind of extension to the modeling work of Lau et al. (2017) would be the most natural response to the variable pattern of performance across evaluation metrics and probe sentences described by Sprouse et al. (2018).

A Manipulation Targeting Latent Factors

In this study we use a simple instruction manipulation to contrast the ratings produced in response to two different questions. One question type asked participants to rate the acceptability of the target sentence, and one asked them to indicate how confident they were that the sentence was grammatical. We ask whether a representative sample of American English speakers would make any distinction at all between these two questions, and if so, what changes in the decision making process might underlie the distinction. The hypothesis that qualitatively different types of judgment might be elicited is suggested by the grammatical illusion literature, to the extent that the striking dissociation between syntactic soundness and acceptability evident in grammatical illusions is thought to be apparent to audiences without extensive training in linguistics. Alternatively, people may not distinguish between the two questions at all, or they may respond with a distinction that has no special relationship with syntactic soundness, such as a uniform reduction in ratings for all sentences in one condition, a move toward more extreme ratings for all sentences, or a change in noise levels. The main goal of this study is to differentiate between these possible scenarios.

To make the contrast between the two question types as salient as possible, we chose a within-subjects design, with each participant giving two blocks of ratings, one for each instruction condition. Items were never rated twice by any one participant. Participants were introduced to the idea of isolating structure from other components of overall acceptability with a brief description of the "colorless green ideas sleep furiously" sentence (Chomsky, 1957)[1] and then asked to rate one block of sentences for overall acceptability and one block for grammatical validity only.

In order to expose the relationship, if any, between the lay interpretation of the two different questions and the distinction drawn between acceptability and structural soundness in linguistics, we presented sentence types commonly described as particularly strong examples of the theoretical dissociation.

One particularly well-known example is center-embedding, which produces sentences widely regarded as grammatical but unacceptable (Chomsky and Miller, 1963; Karlsson, 2007). There are also ungrammatical sentences with unusually high acceptability. Possibly the most well-known is the comparison illusion (Phillips et al., 2011), often illustrated with the example "More people have been to Russia than I have." This sentence is considered unparsable because it has no possible interpretation, and cannot be considered either true or false in any possible state of the world. However, it is generally considered to be more acceptable than might be expected of a nonsense sentence and given the status of a "grammatical illusion." Other phenomena thought to introduce acceptability differences between sentences with equivalent grammatical status include negative polarity item (NPI) illusions (Drenhaus et al., 2005) and agreement attraction sentences (Bock and Miller, 1991). In addition to stimuli constructed to replicate these phenomena, we also examined a set of stimuli drawn from those used in Sprouse et al. (2013) for which expert judgment apparently differed from crowd-sourced judgments, hypothesizing that the difference may have been because different judgment types were applied. The full set of stimuli used are given in **Appendix B**.

State Trace Analysis

To interpret the impact of the instruction manipulation we turn to state-trace analysis (Bamber, 1979; Kalish et al., 2016). State-trace analysis is a tool for identifying dissociable sub-systems in task performance. The "state-trace" at the heart of this analysis is a plot of the co-variation of two dependent variables across different experimental conditions (see Newell and Dunn, 2008 for a review, Dunn and Kalish, 2018 for a more complete treatment). Mathematically, a state trace is a generalization of the yes-no receiver-operating-characteristic curve, a standard tool for evaluating classification accuracy that describes the full range of possible trade-offs between sensitivity and specificity (Bamber, 1979). Under relatively weak assumptions, a state trace plot can be diagnostic of the dimensionality of the underlying process. Single process or single resource accounts, by definition, claim that all possible pairs of outcomes can be described as a point on a single underlying dimension. In this case, points on the state-trace plot are restricted to fall on a one-dimensional line. In contrast, if there are multiple processes or mental resources underlying task performance, points on the state trace plot are not so constrained, and are overwhelmingly more likely to "break the line" than not. Various frequentist (Kalish et al., 2016) and Bayesian (Prince et al., 2012; Davis-Stober et al., 2016; Cox and Kalish, 2019) formulations for state trace analyses exist, but in essence all report on whether or not it is possible to conclude that the "line has been broken" while allowing for noisy measurement. The implementation used here is the frequentist one due to Kalish et al. (2016). This test takes the one-dimensional scenario as the null hypothesis and produces a p-value representing how extreme the observed data are in a bootstrapped population of simulated outcomes drawn from the null. More detail regarding the bootstrap procedure underlying the p-value reported can be found in Wagenmakers et al. (2004). The state trace analysis described in Kalish et al. (2016) uses the data-informed variant of this procedure. A sampling distribution over the difference in fit for the one-dimensional and two-dimensional models is generated using bootstrapped samples drawn from the full

[1]While the colorless green ideas sentence was originally presented as a dissociation between surface probability and structural soundness, modern approaches to language modeling generally agree that the grammatical permutation of these words is indeed more probable than the alternatives (Abney, 1996; Pereira, 2000; Manning, 2002). Accepting this caveat, the sentence remains a striking example of a dissociation between structural soundness and plausibility, and we considered it a good vehicle for communicating the basic idea of isolating judgments of grammatical structure to participants.

data set. At each iteration, goodness of fit values are calculated using a coupled monotonic regression as described by Burdakov et al. (2012). The *p*-value is the proportion of goodness-of-fit differences observed with bootstrapped samples that exceed the difference observed in the full sample. Following the normal logic of *p*-values, if this proportion is large, the observed result is unremarkable under the null hypothesis, while if it is small the observed data constitute an extreme observation if the one-dimensional account were true. Full implementation details appear in Kalish et al. (2016). The motivation and foundations of the procedure are discussed further in Dunn and Kalish (2018). An accessible discussion of applications in psychology appears in Newell and Dunn (2008).

Although in principle any set of data points can be fit by some sufficiently complex one-dimensional line, because the experimental conditions are under the researcher's control they can be selected in order to produce a monotonic relationship with the outcome variables. This assumption of monotonicity has to be defended on its own merits for each application (Ashby, 2014), but if it can be assumed it imposes a severe constraint on the state trace plot. Under these conditions, single process accounts commit to predicting a monotonic state trace. Multiple process accounts can in principle also produce monotonic state traces, but their extra degrees of freedom allow for so many alternative possibilities that the one dimensional result is relatively unlikely and constitutes an extreme observation (for a frequentist) or a highly suspicious coincidence (to a Bayesian).

To illustrate the application of state trace analysis, **Figure 1** presents the results of a series of analyses on simulation data. These simulations were generated by taking the actual experimental data (described below) and substituting simulated responses for the observed ones. In these simulations, each sentence type had associated with it two latent properties specifying the true consensus mean for type-1 ratings and a type-2 ratings. The property underlying type-1 ratings was drawn from a uniform distribution between 0 and 5 (the range of the rating scale used in the study), the property underlying type-2 ratings added gaussian noise to produce a random variable linked to the type-1 property by a specified level of correlation. Observed ratings of each type were generated by adding gaussian noise with standard deviation 1 to the true consensus rating for each sentence. As the correlation between the two latent properties approaches 1, the simulation behaves more and more like a one-dimensional process, the state trace plot produces a thinner one-dimensional line, and the *p*-values associated with the state trace analysis approach a uniform distribution. Conversely, as the correlation between the two latent properties decreases the state trace plot produces a fatter two-dimensional ribbon shape, and the *p*-values associated with the state-trace analysis tend to be confined to low values. A critical advantage of using a state trace analysis over simply examining the correlation between the two rating types is that the state trace analysis is not constrained by an assumption of linearity. By considering only the rank order of items under each rating, it tests for the monotonicity rather than the linearity of a relationship, and makes no distributional assumptions. Since the simulation pictured in **Figure 1** does not vary the form of the relationship between simulated rating type-1 and rating type-2, the distribution of simulated *p*-values in this plot may not reflect the true power of the experimental design. We present it here as a reference for readers who may be unfamiliar with state trace plots.

State trace analyses have been applied effectively across a number of different domains in cognitive science, where questions about the number of processes underlying a phenomenon are common. Example applications include in memory (Dunn, 2008), face recognition (Prince and Heathcote, 2009), and reasoning (Stephens et al., 2018).

Summary

This study presents a state trace analysis of judgment data collected under an instruction manipulation contrasting judgments about sentence structure specifically with judgments about overall sentence acceptability. We use a within-subjects design, and collected probe sentences thought to maximize the distinction between structural and other contributors to overall rating judgments. We consider detailed instructions as complementary to the qualitative review of ratings advocated by Schütze (in press), and the results of this study give a sense of the order of magnitude of instruction-driven effects. By applying a state-trace analysis, we are also able to comment on the interpretation of the different evaluation metrics described by Sprouse et al. (2018) and their implications for language modeling. To foreshadow the results, we do observe differences in ratings for at least some probe sentences that align with the way the linguistics literature typically separates structural features from overall acceptability. Specifically, we find that errors of agreement attraction are rated more leniently for acceptability than grammatical soundness, and center-embedding sentences are rated somewhat more leniently for grammatical soundness than acceptability (although overall ratings for this sentence type are consistently low). Since these differences are in opposite directions, they cannot be accounted for by a simple scaling relation. However, a state-trace analysis of the relationship between the two rating types suggests that they are plausibly monotonically related, leaving open the possibility that a single well-formedness factor underlies both kinds of ratings.

METHOD

The experiment presented instructions asking each participant to rate sentences in two distinct question blocks, one asking about acceptability and the other grammatical soundness. The order of question types was randomized. Each block contained 30 test items. The grammatical soundness block contained two additional attention check items which were excluded from analysis. We turned to the literature on grammatical illusions (Phillips et al., 2011) to find sentence types known to produce striking contrasts between their acceptability and grammatical status. We examined doubly center-embedded relative clauses (Chomsky and Miller, 1963), NPI illusions (Drenhaus et al., 2005), agreement attraction sentences (Bock and Miller, 1991), and comparative illusions (Townsend and Bever, 2001). In addition, we also examined a set of stimuli drawn from those used in Sprouse et al. (2013) for which expert judgment apparently differed from crowd-sourced judgments, hypothesizing that the

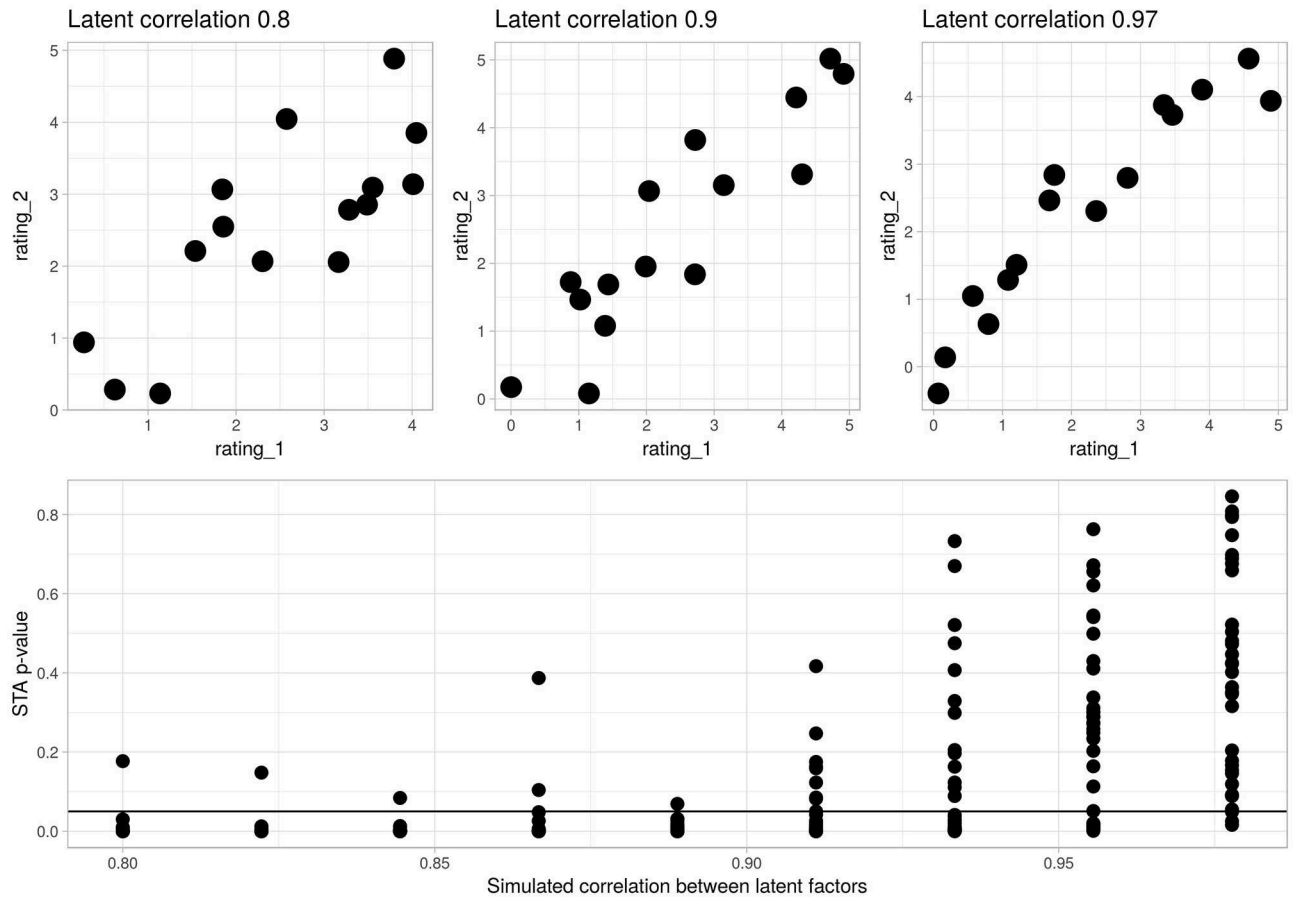


FIGURE 1 | Illustrative state-trace analyses on simulation data. The lower panel presents *p*-values from 315 simulated experiments where ratings type 1 and 2 were correlated to the degree specified on the x-axis. The horizontal bar is at $p = 0.05$. There are 35 repetitions at each degree of correlation. The upper panel presents three representative state-trace plots from simulations run at 0.8, 0.9, and 0.977. The state trace analysis tends to reject the null hypothesis of a single latent factor when the simulated factors are correlated at <0.9, and often fails to reject above that level. It is important to note that the state trace analysis is not just a test of correlation: by considering the consistency of item ranks it avoids assuming linearity.

difference may have been because different judgment types were applied. The full set of stimuli used are given in **Appendix B**.

Stimuli

There were 433 sentences included in the stimuli pool. Of these, 112 were center-embedding sentences, 48 based on stimuli used in Gibson and Thomas (1999), and 64 from Vasishth et al. (2010). Each center-embedding sentence had four variations, one grammatical full version and three ungrammatical partial versions derived by deleting either the first, second, or third verb phrase. An example is "The ancient manuscript that the grad student who the new card catalog had confused a great deal was studying in the library was missing a page." from which "had confused," "was studying," or "was missing a page" can be deleted to create a set of four related sentences. We anticipated that the grammatical full center-embedding would be considered relatively low on acceptability. This sentence structure also shows acceptability differences among the ungrammatical variations. In English, deleting the second verb phrase can improve ratings for this sentence type (Vasishth et al., 2010).

There were 124 agreement attraction sentences, all based on prompts used in Bock and Miller (1991). Each agreement attraction sentence had four variations, grammatical singular-singular agreement, ungrammatical singular-plural clashes, ungrammatical plural-singular clashes, and grammatical plural-plural agreement. An example is "The slogan on the poster is offensive to vegetarians," which with the variations slogan/posters, slogans/poster, and slogans/posters creates a set of four sentences. Although errors are relatively rare in natural language use, agreement attraction errors are among the more common types appearing in English (Bock, 2011) and were anticipated to receive high acceptability ratings alongside low grammaticality ratings.

There were 69 NPI sentences. These were original stimuli intended to follow the NPI illusory licensing pattern (Drenhaus et al., 2005) with reference to example sentences described in Xiang et al. (2009). Each NPI sentence was given in

grammatical "valid licensing" and ungrammatical "partial match" and "unlicensed" forms. One example is "No restaurants that local newspapers have recommended in their dining reviews have ever gone out of business" (valid). "The restaurants that no local newspapers have recommended in their dining reviews have ever gone out of business" (partial match). "Most restaurants the local newspapers have recommended in their dining reviews have ever gone out of business" (unlicensed). The partial match and unlicensed forms were anticipated to give different acceptability ratings despite similar (poor) grammaticality status.

There were 48 comparison illusion sentences, drawn from those used by Wellwood et al. (2018). Each sentence had two variations, one grammatical with compatible comparisons and one ungrammatical illusion sentence with incompatible comparisons. One example is the pair of sentences "Last summer more famous bands had a big stadium show than lesser-known bands did." and "Last summer more famous bands had a big stadium show than the lesser-known band did." Although the form with incompatible comparisons is ungrammatical and admits no possible interpretation, these sentences are often considered to have strikingly high acceptability.

Finally, there were 80 sentences drawn from stimuli used in Sprouse et al. (2013). This study compared the status assigned to sentences by contributors to the journal Linguistic Inquiry (conforming or non-conforming to a particular linguistic pattern under discussion) with acceptability ratings given by naive participants. Although strong agreement was the rule across the majority of items, the sentences used here were drawn from the minority of items for which there was disagreement, i.e., non-conforming items that received above-median acceptability ratings (20 sentences) or conforming items that received below-median acceptability ratings (60 sentences). Unlike the other sentence types, these sentences were heterogeneous in structure. One example of a poorly-rated but pattern-conforming sentence is "We proved Susan to the authorities to be the thief." One non-conforming but highly-rated sentence is "All the postal workers seem to have all taken a break at the same time." Unlike the other stimuli considered here, none of these sentences have been claimed to produce "illusions" directly dissociating the acceptability and grammatical status of any single item. However, we considered it possible that the apparent conflict between the pattern conforming/violating status of these examples and their crowd-sourced acceptability scores is that the two reflect different kinds of judgment.

Presentation

Stimuli were presented to participants as a web page. The landing page contained a consent preamble, after which participants were given instructions describing the two kinds of judgments they would be asked to make. The instructions are given in full in **Appendix A**. Grammaticality judgments were described as rating the participant's confidence in whether or not an item "follows the rules" for constructing an English sentence. Participants were asked to "label all sentences with a grammatical error as ungrammatical, even if the error is small, and label all sentences with no errors as grammatical, even if they are badly written or unclear." Acceptability was described as a broader concept "more about how natural a sentence sounds." with the explanation that "Among all grammatical sentences, some will be highly acceptable and 'sound good' while others will be not very acceptable and 'sound bad,' even though they're all grammatical. Similarly, although ungrammatical sentences tend to 'sound bad,' some are worse than others." Participants needed to pass a comprehension quiz to progress from the instructions to the study task. This consisted of a two-item multiple-choice quiz that asked "For this study, which of these best describes a **grammatical** sentence?" with the expected answer "A sentence that 'follows the rules' of English, whether it makes sense or not." and "For this study, which of these best describes an **acceptable** sentence?" with the expected answer "A sentence that 'sounds good,' or 'sounds natural'." Multiple attempts were allowed, however failed attempts returned participants to the beginning of the instruction sequence. Before continuing to the study task, participants were asked to self-report age, gender, native language, and current country of residence. Each question block was preceded by a short prompt screen recalling the instructions. For the grammaticality judgement block, the prompt screen read "This block of questions asks you to judge if a sentence is grammatical or not. It doesn't matter if the sentence is ugly or even makes no sense: please answer 'Yes' if it is a valid construction in English or 'No' if it is not." For the acceptability judgment block, the prompt screen read "This block of questions asks you to judge how acceptable a sentence is. Here 'acceptable' means 'well-formed' or 'natural sounding.' The sentences here range in acceptability from very good to very poor, please use the rating scale to indicate where each sentence falls in this range." The order of blocks was randomized. Trials displayed a html h1 title with the current question type, either "Is this a valid grammatical English sentence?" or "Is this an acceptable English sentence?" Centered under the title was a box with a 1px solid green border containing the test sentence. The dimensions of this box may have varied depending on participant's device and browser, font size was 1.5 em. Response options were displayed under the test item outside this bounding box, and consisted of an evenly spaced row of six html buttons. In the grammaticality judgment block these were labeled "Definitely not grammatical," "Probably not grammatical," "Possibly not grammatical," "Possibly grammatical," "Probably grammatical," and "Definitely grammatical." In the acceptability judgment block they were labeled "Highly unacceptable," "Unacceptable," "Somewhat unacceptable," "Somewhat acceptable," "Acceptable," and "Highly acceptable." This labeling for the response options does introduce a difference between instruction conditions, in that the "grammaticality" judgment is presented as a rating of confidence while the "acceptability" judgment is presented as one of degree. This design choice allowed us to describe "grammaticality" to participants as a categorical structural property without varying the number of responses available.

Participants progressed to the next item immediately on making each response. The response buttons were disabled for the first 1,000 ms of each trial. Each of the two blocks consisted of 30 probe items randomly drawn from the pool of stimuli, with the "grammatical" judgement block also containing two additional attention check items. Item draws were without

replacement so participants never viewed the same sentence twice. Randomization was uniform over sentence type, the main unit of analysis, rather than uniform over items. The attention check items added to the "grammatical" judgment block were identical for every participant: "Sarah expected to get a good grade." and "Him would have been fired." These were considered to have known grammatical status (high and low, respectively) and were used as attention checks, triggering the exclusion of participants who gave an unexpected rating to either item. These attention check items were not included in the main analysis.

Participants

324 participants were recruited via Amazon Mechanical Turk. Ages ranged from 18 to 71, with a mean age of 35, 126 female and 4 declining to give a gender. We interpret these responses as indicating our sample is an typical of the Mechanical Turk workplace, but note that as well as being extremely WEIRD in the sense of Henrich et al. (2010), the MTurk worker pool has a slightly higher average education than the US population at large (Levay et al., 2016), which may be relevant to the interpretation of the instruction manipulation. A relatively large proportion of recruited participants were excluded from analysis. Five reported being non-native English speakers, 26 did not complete all questions, 61 completed with unrealistically fast response times (defined as <4 min), and 108 gave unexpected answers to the attention check questions, either failing to use one of the lowest two response options for "Him would have been fired" or failing to use one of the highest two response options for "Sarah expected to get a good grade." Fifty-five participants triggered multiple exclusion criteria, in total 135 recruited participants were excluded and 189 retained (58% retention). Mean total participation time was ∼10 min, including time spent reading the instructions. Mean response time per-item was 7.7 s.

Results

On average each sentence received 13 ratings in each instruction condition, the first and third quartiles were 8 and 18 ratings per item, respectively. Ratings were coded between 0 and 5.

Response to Instructions

A natural first question is whether the instruction manipulation produced any difference in responding at all.

Responses to the grammaticality question were both slower and more extreme than responses to the acceptability question. After standardizing response times for each participant, responses to grammaticality questions were on average 0.1 standard deviations slower than participants' overall mean response time, while responses to acceptability questions were 0.01 standard deviations faster. This difference was statistically significant ($t_{df=11231} = -6.62, p < 0.001$). Responses to the grammaticality question were also numerically more extreme, as shown in **Figure 2**. Only 31% of responses to the "acceptability" question used the most extreme options in either direction, compared to 48% of responses to the "grammatical" question.

The sentence types used in this stimulus set were chosen to give the best possible chance of dissociating ratings emphasizing structural well-formedness from those based on overall acceptability as described in the instructions. Agreement attraction sentences were hypothesized to be highly acceptable, even in their ungrammatical variations. Center-embedding sentences were predicted to be rated as grammatical but unacceptable, and the missing VP2 variation was expected to receive more favorable acceptability ratings while having the same grammatical status as the other missing verb-phrase variants. Comparison illusion sentences were expected to be higher in acceptability ratings than grammaticality ratings for the incompatible-comparison variation only. NPI illusion sentences were expected to be rated as more acceptable under partial match than unlicensed variations, with both having low grammaticality ratings. The pattern-conforming and non-conforming examples from Linguistic Inquiry articles were expected to receive ratings more in line with their pattern-conforming status when rated under grammaticality instructions than acceptability instructions. **Figure 3** summarizes the differences found graphically, with the corresponding mean ratings and *t*-tests for the difference between means given in **Table 1**.

It is clear that the characterization of sentences as acceptable-but-not-grammatical or grammatical-but-not-acceptable is not reflected in an absolute sense in these ratings. However, in at least some cases participants appeared to distinguish between the instruction sets selectively for particular sentence types. Agreement attraction sentences were rated more leniently under acceptability instructions than grammaticality instructions, while center-embedding sentences showed the reverse pattern.

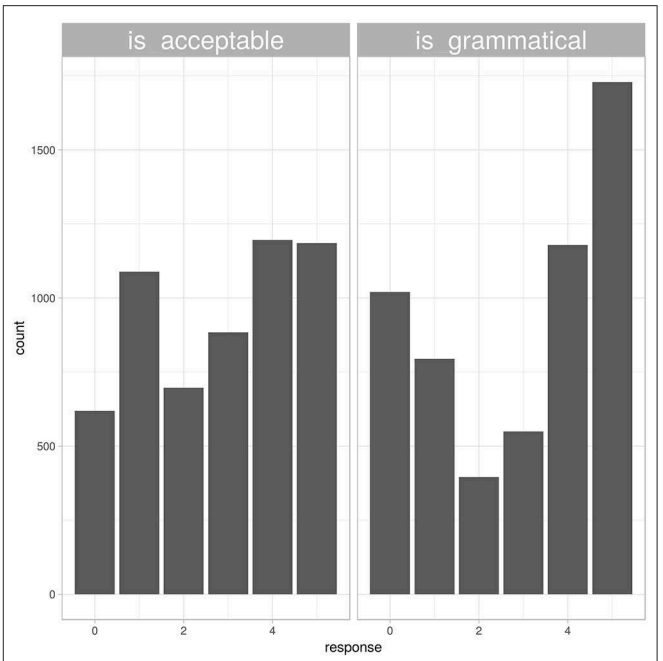


FIGURE 2 | Histogram of responses from different instruction blocks. Responses to the question "is this an acceptable sentence" (left panel) were more uniformly distributed than responses to the question "is this a grammatical sentence" (right panel).

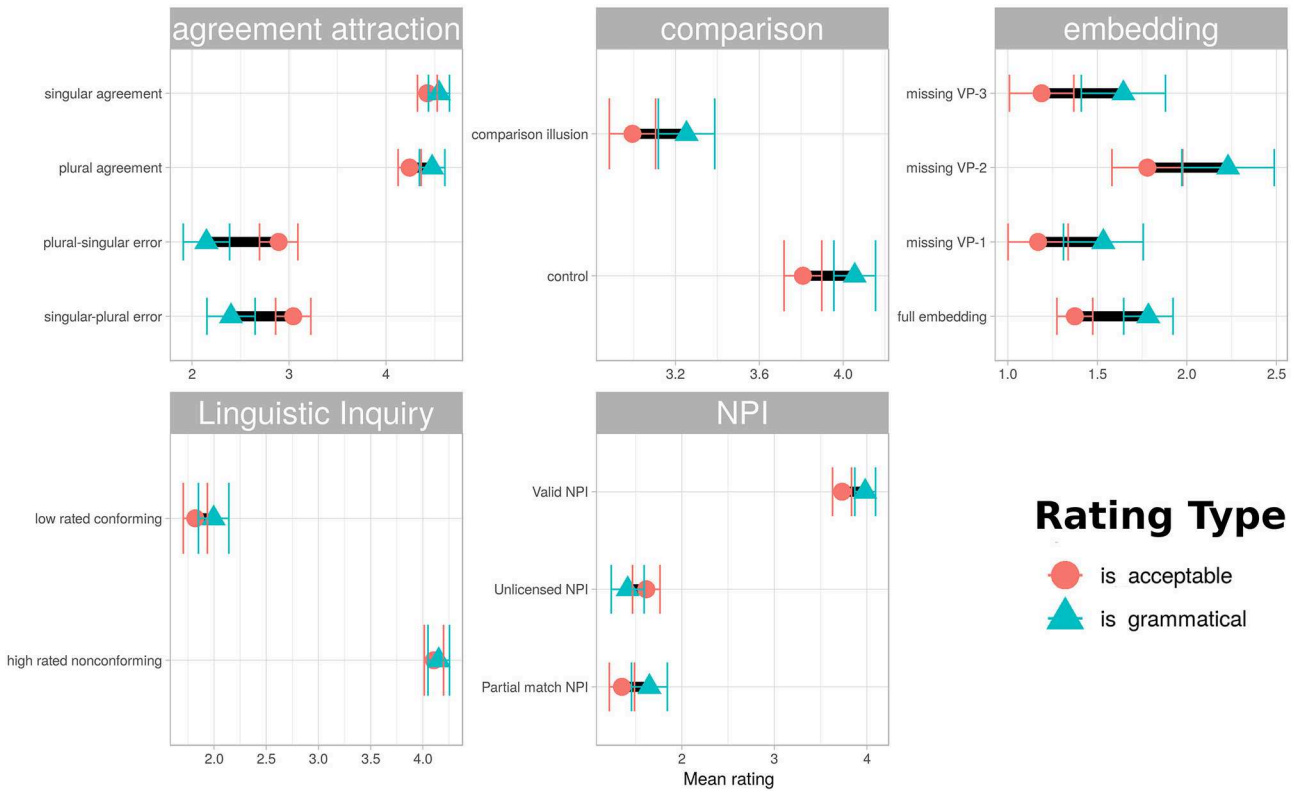


FIGURE 3 | Mean ratings for various sentence types under the two instruction conditions. Intervals are 95% confidence. Although ratings under the two different instruction types were strongly correlated, the instructions did appear to produce relative differences, with some sentences receive more lenient ratings under one or the other instruction condition. Agreement attraction errors were rated more leniently under "acceptability" instructions than "grammaticality" instructions, while center-embedding sentences showed the reverse pattern, being more leniently rated under "grammaticality" instructions. Comparison sentences were more leniently rated under "grammaticality" instructions, but the difference between the two rating types was similar for both illusion and control sentences. The NPI items and those drawn from Linguistic Inquiry articles received similar ratings under both instructions.

Figure 4 shows an alternative visualization of the differences, plotting the distribution of rating differences for individual items under different instruction conditions, grouped by sentence structure type. This view emphasizes the difference in variability masked by the more conventional comparison of mean ratings in **Figure 3**. Sentences canonically regarded as grammatical show markedly less variability than their ungrammatical counterparts.

State Trace Analysis

The particular implementation of state trace analysis used here is due to Kalish et al. (2016)[2]. In brief, the test examines the rank ordering of the stimuli under both rating types. If the rank orderings are consistent, the two rating outcomes are monotonically related and the state-trace plot is one-dimensional (although not necessarily linear). Otherwise they are not monotonically related, and the state trace plot is two-dimensional. With real-world experimental data, some sampling noise is expected, such that "minor" violations of rank ordering need not necessarily imply the two-dimensional outcome has been obtained. The implementation described by Kalish et al. (2016) takes the one-dimensional result as the null hypothesis, and determines via a non-parametric bootstrap procedure how extreme the observed violation is relative to those found in a bootstrapped population of possible results under the null hypothesis. This quantity is a p-value and admits the usual interpretation, with rejection of the null corresponding to a conclusion that the state-trace is two dimensional with a type-I error rate determined by the chosen alpha. No assumptions regarding the data-generating distributions are required. There is a minimum number of data points to avoid degenerate re-sampling in the bootstrap (found in simulation to be $n \approx 8$), a requirement that is met by this data set.

The eponymous state-trace plot is given in **Figure 5**. On inspection, the points appear to lie on a one-dimensional S-shaped curve. The p-value associated with this configuration is $p = 0.18$, failing to reject the one-dimensional null hypothesis.

Discussion

Given this pattern of responses, what can be concluded about the impact of the instruction differences? How does this contribute to a more complete description of rating

[2]Using the implementation https://github.com/michaelkalish/STA (accessed February 25, 2019).

TABLE 1 | Means and significance tests corresponding to **Figure 3**.

Phenomenon	Item type	Mean acceptability	Mean grammaticality	*t*-test
Linguistic Inquiry	Low rated conforming	1.8	2	No difference
Linguistic Inquiry	High rated non-conforming	4.1	4.2	No difference
Embedding	Full embedding	1.4	1.8	$t_{df=1000} = -4.7$,p = 2.6e−06
Embedding	Missing VP-2	1.8	2.2	$t_{df=360} = -2.7, p = 0.0069$
Embedding	Missing VP-3	1.2	1.6	$t_{df=350} = -3, p = 0.0027$
Embedding	Missing VP-1	1.2	1.5	$t_{df=380} = -2.6, p = 0.011$
Agreement attraction	Plural-singular error	2.9	2.1	$t_{df=550}$ =4.7,p = 3.2e−06
Agreement attraction	Plural agreement	4.2	4.5	$t_{df=550} = -2.6, p = 0.01$
Agreement attraction	Singular agreement	4.4	4.5	No difference
Agreement attraction	Singular-plural error	3	2.4	$t_{df=520} = 4.1$,p = 5e−05
NPI	Valid NPI	3.7	4	$t_{df=1100} = -3.2, p = 0.0013$
NPI	Partial match NPI	1.4	1.6	$t_{df=490} = -2.5, p = 0.014$
NPI	Unlicensed NPI	1.6	1.4	No difference
Comparison	Control	3.8	4.1	$t_{df=1100} = -3.6, p = 0.00032$
Comparison	Comparison illusion	3	3.3	$t_{df=1100} = -2.9, p = 0.0038$

Most differences between rating types are significant at $p < 0.01$ level. One exception was the items drawn from examples used in Linguistic Inquiry articles, which received indistinguishable ratings under both instructions.

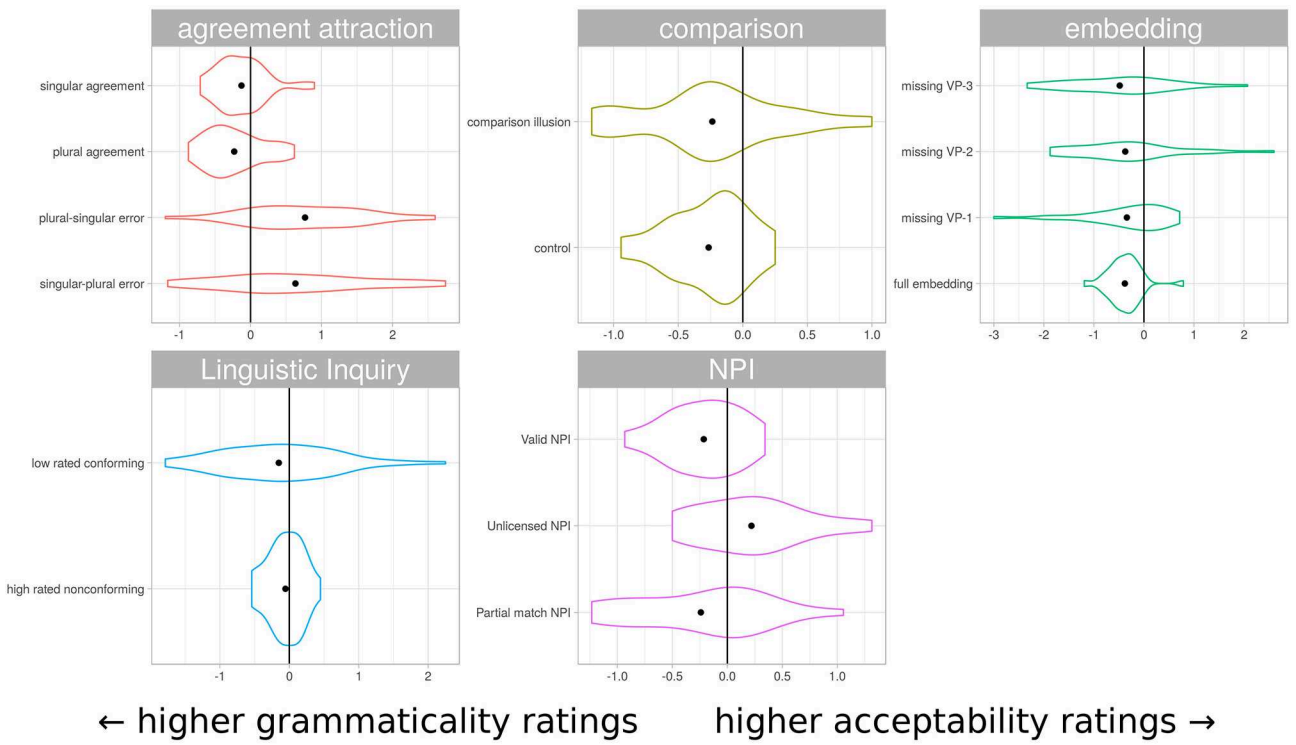


FIGURE 4 | Distribution of rating differences under different instruction sets. Each violin represents the distribution of mean item rating difference under the different instruction conditions. All differences are *acceptabilityrating − grammaticalityrating*, so items rated more favorably under grammaticality instructions tend to the left. This visualization represents the same data as **Figure 3**, but obscures absolute rating magnitude and displays richer information about the distribution of differences. The sentence classes canonically regarded as grammatical typically have much lower variability in difference scores.

task behavior, and in particular the potential disconnect between what linguists want to know and how participants do the task?

First, the results suggest that people are quite sensitive to the wording of rating task instructions. Ratings for the same sentences differed across the two question blocks as a

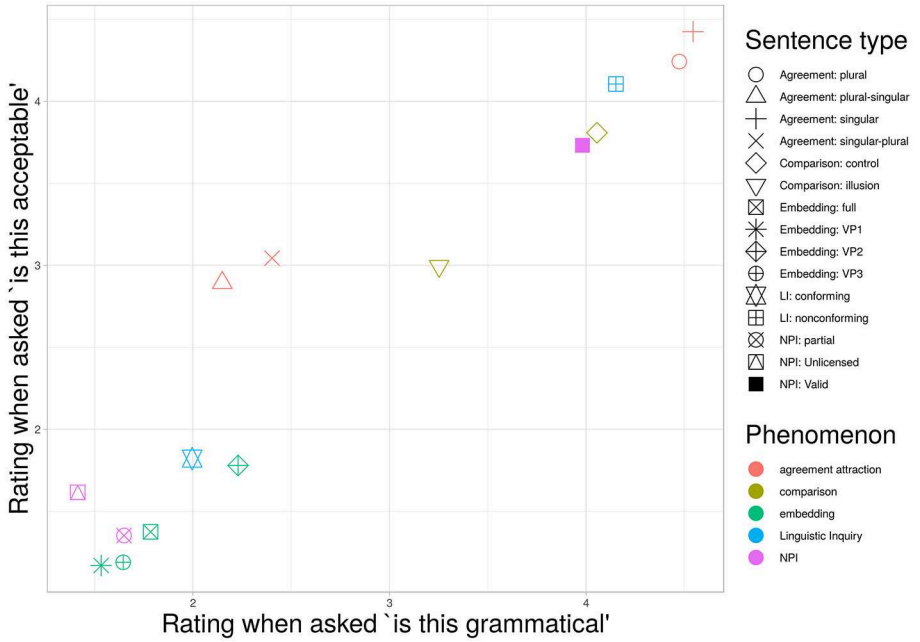


FIGURE 5 | State trace plot. The state-trace analysis used here asks if the rank ordering of two outcome variables is consistent, the visual signature of which is a monotonic relationship when one is plotted against the other. Here, the ratings under the two different instruction conditions do appear monotonically related, although interpreted as such the relationship is not linear.

result of instruction wording. Not only were responses in the "grammatical" question block more extreme (which might also be expected from a simple demand effect), people's ratings differentiated between at least two of the probe phenomena in the opposing directions. Errors of agreement attraction were rated more leniently under acceptability instructions and center-embedding sentences were rated more leniently under grammaticality instructions. The impact of the instructions cannot be a simple scalar shift in lenience or a change in the volatility of ratings. However, despite producing differences with opposite signs, the ratings across the two instruction conditions remained consistent with a one-dimensional state trace. The relevance of the response to the instruction manipulation to the theoretical distinction between grammaticality and acceptability lies not in any direct mapping between the two but rather by requiring accounts of the rating process to accommodate both the differential impact of the instruction manipulation on different sentence types while maintaining a single underlying dimension on which they vary.

The signal detection model of rating behavior proposed by Bader and Häussler (2010) is one such account. Under this description of the rating task, the impact of the instruction manipulation could be described as a reduction of noise and an increase in caution when rating under "grammaticality" instructions relative to "acceptability" instructions. The slight increase in response times observed for the "grammaticality" questions is also consistent with this interpretation. The viability of the signal detection model undermines arguments that grammatical illusion phenomena demonstrate the need to appeal to multiple qualitatively different factors in lay ratings of sentences. It is not the case that dissociations between grammatical status and acceptability rating are only produced by experts working under a technical definition of grammaticality: naive participants in this study also sometimes produce such dissociations for lay interpretations of grammaticality. It remains possible given these data that the two are related, and that the limiting tendency of an ideal acceptability judgment under noiseless conditions and high caution may potentially reproduce the expert pattern without necessarily invoking distinct latent components of ratings. Further, "expert-like" and "acceptability-like" patterns may be apparent to the same people at the same time, as they were to the participants in this study, if the underlying goodness quality is interrogated in different ways, as occasioned in this example by simply changing the wording of the question.

Arguments from the subjective experience of dissociating judgments in grammatical illusion phenomena are not the only evidence for a separation between latent components of acceptability ratings. Arguments highlighting systematic deficits in the performance of language models trained to predict acceptability judgement without recourse to an explicitly separate syntactic information (Dyer et al., 2016; Sprouse et al., 2018; Warstadt and Bowman, 2019; Warstadt et al., 2019) suggest endorsing the "component" interpretation. However, whether these deficits are inherent to all such approaches or simply reflect the peculiarities of current state of the art is an open question. The effect of the instruction manipulation presented here on ratings suggests that an alternative argument from grammatical

illusion phenomena is unsound: it is possible that subjective dissociations between canonical status and acceptability rating can be accommodated by appealing to *unbiased* noise and caution only. We do not claim that participants were producing judgments of grammaticality in the technical sense when asked to "rate for grammaticality," but we observe that whatever the change in the decision making process was, it moved rating judgments toward the outcomes that would be expected from expert analysis and did so by a mechanism consistent with increased deliberation only.

The results are subject to a number of caveats. Most importantly, the one-dimensional state trace result is subject to the usual cautions against interpreting a failure to reject as evidence for the null: it may be that the fifteen sentence types represented in this study are simply not diagnostic. The phenomena used here were chosen to maximize the chance of finding a dissociation between structural and other features, but it is definitely possible that a broader survey of the stimulus space would identify a dissociation where these sentences did not.

The results only apply to the particular population sampled. Because the outcomes rely heavily on the culturally-bound interpretation of the instructions, this study is tightly constrained by the limitations of WEIRD participant pools (Henrich et al., 2010), such as Mechanical Turk workers (Levay et al., 2016). The results of the instruction manipulation may depend on education level, age cohort, or handedness, and the analyses presented here provide no mechanism for identifying systematic effects due to such factors or mitigating them if found. In particular, any differences in ratings due to education up to and including linguistic-specific expertise would be highly desirable, but these data do not support such an analysis.

It's not clear which elements of the instruction manipulation were responsible for producing the differences in ratings. Candidate elements include the description of the judgment types, the colorless green ideas example, the attention check quiz associated with the instructions, and the labeling of the response options. In particular, the decision to express "grammaticality judgments" as a rating of confidence in the presence or absence of errors may have encouraged a different pattern of results to that which would have been obtained under some alternative set of response options. We considered matching the numerical range of the two rating scales to be the conservative choice when testing for differences between them.

One motivation for this work was to quantify the extent to which detailed instructions can help control variability in the motivation behind ratings identified by Schütze (in press). Although the instructions were quite brief and participants on the Mechanical Turk platform are often highly motivated to finish studies quickly, we find statistically significant differences in ratings due to the instruction manipulation. Although these data argue against appealing to people's ability to isolate any specific component of overall acceptability, they also show that rating tasks drawing attention to structural components of acceptability specifically can produce qualitative differences in ratings that may be meaningful on the scale of typical effect sizes in linguistics. The main contribution of this study is to point out that there is no contradiction between these two things, they can both be true at the same time.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Michigan Health Sciences and Behavioral Sciences review board, irbhsbsumich.edu. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SL designed the study, collected the data, and drafted the initial write-up under the supervision of RL, who contributed to all stages of this process. RS and JD supplied the code and documentation used in the analysis, reviewed the analysis, and contributed significantly to editing the initial draft for clarity and correctness.

FUNDING

This work was supported by the College of Literature, Science and the Arts, University of Michigan.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02886/full#supplementary-material

REFERENCES

Abney, S. (1996). "Statistical methods and linguistics," in *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, eds J. Klavans and P. Resnik (Cambridge, MA: MIT Press), 1–26.

Adger, D. (2003). *Core Syntax: A Minimalist Approach*, Vol. 20. Oxford: Oxford University Press.

Ahler, D. J., Roush, C. E., and Sood, G. (2019). "The micro-task market for lemons: data quality on Amazon's Mechanical Turk," in *Meeting of the Midwest Political Science Association* (Chicago, IL).

Ashby, F. G. (2014). Is state-trace analysis an appropriate tool for assessing the number of cognitive systems? *Psychon. Bull. Rev.* 21, 935–946. doi: 10.3758/s13423-013-0578-x

Bader, M., and Häussler, J. (2010). Toward a model of grammaticality judgments. *J. Linguist.* 46, 273–330. doi: 10.1017/S0022226709990260

Bamber, D. (1979). State-trace analysis: a method of testing simple theories of causation. *J. Math. Psychol.* 19, 137–181.

Bock, K. (2011). How much correction of syntactic errors are there, anyway? *Lang. Linguist. Compass* 5, 322–335. doi: 10.1111/j.1749-818X.2011.00283.x

Bock, K., and Miller, C. A. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93.

Braze, F. D. (2002). *Grammaticality, acceptability and sentence processing: a psycholinguistic study* (Ph.D. thesis), University of Connecticut, Storrs, CT, United States.

Burdakov, O., Dunn, J., and Kalish, M. (2012). "An approach to solving decomposable optimization problems with coupling constraints," in *21st International Symposium on Mathematical Programming* (Berlin), 271.

Chater, N., and Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends Cogn. Sci.* 10, 335–344. doi: 10.1016/j.tics.2006.05.006

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton & Co.

Chomsky, N., and Miller, G. (1963). "Introduction to the formal analysis of natural languages," in *Handbook of Mathematical Psychology: I.* Vol. 2, eds R. Luce, R. R. Bush, and E. E. Galanter (New York, NY: John Wiley), 269–321.

Cox, G. E., and Kalish, M. L. (2019). Dial M for monotonic: a Kernel-based bayesian approach to state-trace analysis. *J. Math. Psychol.* 90, 100–117. doi: 10.1016/j.jmp.2019.02.002

Davis-Stober, C. P., Morey, R. D., Gretton, M., and Heathcote, A. (2016). Bayes factors for state-trace analysis. *J. Math. Psychol.* 72, 116–129. doi: 10.1016/j.jmp.2015.08.004

Domaneschi, F., and Di Paola, S. (2018). The processing costs of presupposition accommodation. *J. Psycholinguist. Res.* 47, 483–503. doi: 10.1007/s10936-017-9534-7

Drenhaus, H. Frisch, S., and Saddy, D. (2005). "Processing negative polarity items: when negation comes through the backdoor," in *Linguistic Evidence-Empirical, Theoretical, and Computational Perspectives*, eds S. Kepser and M. Reis (Berlin: Mouton de Gruyter), 145–165.

Dunn, J. C. (2008). The dimensionality of the remember-know task: a state-trace analysis. *Psychol. Rev.* 115:426. doi: 10.1037/0033-295X.115.2.426

Dunn, J. C., and Kalish, M. L. (2018). *State-Trace Analysis*. New York, NY: Springer.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). "Recurrent neural network grammars," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, CA), 199–209.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv [Preprint]. arXiv 1811.12231*.

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cogn. Process.* 14, 225–248.

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*, Vol. 1. New York, NY: Wiley.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83. doi: 10.1017/S0140525X0999152X

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., and Snider, N. (2013). The source ambiguity problem: distinguishing the effects of grammar and processing on acceptability judgments. *Lang. Cogn. Process.* 28, 48–87. doi: 10.1080/01690965.2011.572401

Kalish, M. L., Dunn, J. C., Burdakov, O. P., and Sysoev, O. (2016). A statistical test of the equality of latent orders. *J. Math. Psychol.* 70, 1–11. doi: 10.1016/j.jmp.2015.10.004

Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *J. Linguist.* 43, 365–392. doi: 10.1017/S0022226707004616

Lappin, S., and Lau, J. H. (2018). *Gradient Probabilistic Models vs. Categorical Grammars: A Reply to Sprouse et al.* Cambridge: Informal communication Ted Gibson.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cogn. Sci.* 41, 1202–1241. doi: 10.1111/cogs.12414

Levay, K. E., Freese, J., and Druckman, J. N. (2016). The demographic and political composition of mechanical turk samples. *Sage Open* 6, 1–17. doi: 10.1177/2158244016636433

Loftus, G. R. (1978). On interpretation of interactions. *Mem. Cogn.* 6, 312–319.

Mahowald, K., Graff, P., Hartman, J., and Gibson, E. (2016). SNAP judgments: a small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language* 92, 619–635. doi: 10.1353/lan.2016.0052

Manning, C. D. (2002). "Probabilistic syntax," in *Probabilistic Linguistics*, eds J. H. Rens Bod and S. Jannedy (Cambridge, MA: MIT Press), 289–341.

Masia, V. (2017). *Sociobiological Bases of Information Structure*, Vol. 9. Amsterdam: John Benjamins Publishing Company.

Myers, J. (2009). Syntactic judgment experiments. *Lang. Linguist. Compass* 3, 406–423. doi: 10.1111/j.1749-818X.2008.00113.x

Newell, B. R., and Dunn, J. C. (2008). Dimensions in data: testing psychological models using state-trace analysis. *Trends Cogn. Sci.* 12, 285–290. doi: 10.1016/j.tics.2008.04.009

Pereira, F. (2000). Formal grammar and information theory: together again? *Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci.* 358, 1239–1253. doi: 10.1098/rsta.2000.0583

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Exp. Interfaces* 37, 147–180. doi: 10.1163/9781780523750_006

Prince, M., Hawkins, G., Love, J., and Heathcote, A. (2012). An R package for state-trace analysis. *Behav. Res. Methods* 44, 644–655. doi: 10.3758/s13428-012-0232-y

Prince, M., and Heathcote, A. (2009). "State-trace analysis of the face-inversion effect," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Amsterdam).

Schütze, C. T. (in press). "Acceptability ratings cannot be taken at face value," in *Linguistic Intuitions*, eds S. Schindler, A. Drozdzowicz, and K. Brøcker (Oxford: Oxford University Press).

Schütze, C. T., and Sprouse, J. (2014). Judgment data. *Res. Methods Linguist.* 27, 27–50. doi: 10.1017/CBO9781139013734.004

Sprouse, J. (2018). "Acceptability judgments and grammaticality, prospects and challenges," in *Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics*, Vol. 129, eds N. Hornstein, H. Lasnik, P. Patel-Grosz, and C. Yang (Berlin: Walter de Gruyter), 195–224.

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua* 134, 219–248. doi: 10.1016/j.lingua.2013.07.002

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., and Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *Linguist. Rev.* 35, 575–599. doi: 10.1515/tlr-2018-0005

Stephens, R., Matzke, D., and Hayes, B. (2019). Disappearing dissociations in experimental psychology: using state-trace analysis to test for multiple processes. *J. Math. Psychol.* 90, 3–22. doi: 10.1016/j.jmp.2018.11.003

Stephens, R. G., Dunn, J. C., and Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychol. Rev.* 125, 218–244. doi: 10.1037/rev0000088

Townsend, D. J., and Bever, T. G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Vasishth, S., Suckow, K., Lewis, R. L., and Kern, S. (2010). Short-term forgetting in sentence comprehension: crosslinguistic evidence from verb-final structures. *Lang. Cogn. Process.* 25, 533–567. doi: 10.1080/01690960903310587

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., and Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *J. Math. Psychol.* 48, 28–50. doi: 10.1016/j.jmp.2003.11.004

Warstadt, A., and Bowman, S. R. (2019). Grammatical analysis of pretrained sentence encoders with acceptability judgments. *arXiv [Preprint]. arXiv 1901.03438*.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Trans. Assoc. Comput. Linguist.* 7, 625–641. doi: 10.1162/tacl_a_00290

Wellwood, A., Pancheva, R., Hacquard, V., and Phillips, C. (2018). The anatomy of a comparative illusion. *J. Semant.* 35, 543–583. doi: 10.1093/jos/ffy014

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest: 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

HYPOTHESIS AND THEORY
published: 30 January 2020
doi: 10.3389/fpsyg.2019.02972

Interpreting Degree Semantics

*Alexis Wellwood**

School of Philosophy, University of Southern California, Los Angeles, CA, United States

Contemporary research in compositional, truth-conditional semantics often takes judgments of the relative unacceptability of certain phrasal combinations as evidence for lexical semantics. For example, observing that *completely full* sounds perfectly natural whereas *completely tall* does not has been used to motivate a distinction whereby the lexical entry for *full* but not for *tall* specifies a scalar endpoint. So far, such inferences seem unobjectionable. In general, however, applying this methodology can lead to dubious conclusions. For example, observing that *slightly bent* is natural but *slightly cheap* is not (that is, not without a "too cheap" interpretation) leads researchers to suggest that the interpretation of *bent* involves a scalar minimum but *cheap* does not, contra intuition—after all, one would think that what is minimally cheap is (just) free. Such claims, found in sufficient abundance, raise the question of how we can support semantic theories that posit properties of entities that those entities appear to lack. This paper argues, using theories of adjectival scale structure as a test case, that the (un)acceptability data recruited in semantic explanations reveals properties of a two-stage system of semantic interpretation that can support divergences between our semantic and metaphysical intuitions.

Keywords: scale structure, truth conditional meaning, semantic anomaly, language and mind, compositional semantics

OPEN ACCESS

Edited by:
M. Teresa Espinal, Autonomous University of Barcelona, Spain

Reviewed by:
Elena Castroviejo, IKERBASQUE Basque Foundation for Science, Spain
Christopher Kennedy, University of Chicago, United States

***Correspondence:**
Alexis Wellwood
wellwood@usc.edu

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

Received: *17 October 2019*
Accepted: *16 December 2019*
Published: *30 January 2020*

Citation:
Wellwood A (2020) Interpreting Degree Semantics. Front. Psychol. 10:2972. doi: 10.3389/fpsyg.2019.02972

1. INTRODUCTION

This paper examines a corner of semantic theory that has received a lot of attention in the recent linguistic and philosophical literature: the recruitment of 'scale structure' in compositional accounts of the interpretation of sentences like (1) and (2)[1].

(1) a. Ann's glass is full.
b. Ann's daughter is tall.

(2) a. Ann's glass is fuller than Bill's is.
b. Ann's daughter is taller than Bill's is.

According to degree-based theorists, we can learn something about the meaning of (2) by thinking about how its parts compositionally determine truth conditions as follows: (2a) is true only if the degree to which Ann's glass is full is greater than the degree to which Bill's glass is full, and (2b) is true only if the degree to which Ann's daughter is tall is greater than the degree to which Bill's daughter is tall. Correspondingly, without specification of an explicit standard for comparison by a phrase like *than Bill's is*, (1a) is true only if the degree to which Ann's glass is full exceeds the contextually given standard for fullness, and (1b) only if where her daughter's height exceeds the relevant contextual standard; and so on.

[1]Degree-based theories have dominated recent discussion of structures like (1) and (2) as well as those of comparative sentences with *more*, *as*, etc. Degree-based theories contrast with delineation-based theories (see Burnett, 2016 for discussion and citations), but I only address the former here.

Degree-semantic theories provide the rudiments for expanding outward the kinds of structures that can be compositionally interpreted with few additional assumptions, and they work to capture the kinds of inferences that we intuitively find to hold between relevant sentences. For instance, the simple appeal to a greater-than relation between degrees correctly predicts that if (2a) is true, then (3a) is false (evaluated in the same contexts) and, in turn, appealing to the intuitive idea that the negative of an antonymic pair reverses the ordering relation, correctly predicts that if (2b) is true so is (3b).

(3) a. Bill's glass is fuller than Ann's glass.
b. Bill's daughter is shorter than Ann's is.

Very quickly, though, a theory that was primarily designed to offer perspicacious compositions that get the truth-value judgments right is leveraged to explain patterns of semantic anomaly. Here are some examples of the kinds of observations I have in mind. First, while it is possible[2] to construct a comparative construction that targets two adjectives simultaneously, (4a), in many cases the result is anomalous, e.g., (4b)[3].

(4) a. The ladder is taller than Ann's son is wide.
b. ? Ann's glass is fuller than the ladder is tall.

Second, while a modifier like *completely* can sometimes be used to indicate maximal extent along a given dimension, e.g., (5a), in many cases it cannot, e.g., (5b).

(5) a. Ann's glass is completely full.
b. ? Ann's daughter is completely tall.

The going explanation for the asymmetry in (4) assumes that it is only possible to evaluate a comparative relation between two degrees if those degrees share a dimension; (4b), then, is anomalous because the scales associated with *tall* and *full* order degrees along different dimensions. The going explanation for (5) relates to the structure of the degrees so ordered: *completely* picks out the topmost, or maximal, degree on a scale; by hypothesis, the scale associated with *full* provides such an element, but that associated with *tall* does not[4].

Let us take a closer look at these explanations. First, we observe that two strings of words that appear to be syntactically equivalent differ in acceptability. Next, we link these differences in acceptability to differences in the sorts of things that the expressions occurring in the sentence are about:[5] *tall* is about length in the vertical dimension while *full* is about something else. Finally, depending on the target observation, different features of those things are recruited to explain the anomaly: *completely* relates to a scalar endpoint, and some dimensions (like vertical distance) apparently lack such points. Research can get off the ground and continue in stride without wondering much about what is meant when we say "what the expressions are about," but, ultimately, we will want to know whether such explanations are correct or not. If it turned out, for example, that *tall* was not actually about vertical distance or that *full* actually was, or if it turned out that the scale of tallness in fact had an upper bound while the scale of fullness did not, that would certainly be problematic for the theory. But how can we tell what these scales are like, independently of the linguistic diagnostics?

The trouble facing such explanations is put into stark relief when we find clear examples where the linguistic tests turn up results that run counter to our intuitions about what there is (what we may call our *metaphysical intuitions*). For example, just as *completely* is thought to be licensed by gradable adjectives that are associated with scalar maxima, it is contended that *slightly* is licensed with gradable adjectives that associate with scalar minima. In this light, consider the asymmetry in (6).

(6) a. The rod is slightly bent.
b. ? The dress is slightly inexpensive.

The explanation for (6) should run as follows: since *bent* is associated with a scalar minimum, *slightly* is licensed in (6a); but since *cheap* does not so associate, *slightly* is not licensed in (6b). But this seems odd: if we would otherwise suppose that the scale of inexpensiveness is, or is isomorphic to, a scale of cost, there should be a minimum element that is simply 0 dollars (or whatever). We have here a mismatch between the acceptability data and our intuitions about what there is; yet the explanation for the former would seem, on extant accounts, to depend on facts about the latter[6].

Moreover, such cases can be multiplied. von Stechow (1984) and Rullmann (1995) suppose that (7) is deviant because *tall* associates with a scale that has no maximum. [This account dovetails, of course, with the expectations of other authors' interpretations of the fact that *completely* is anomalous with *tall*, recall (5b)].

[2]In English at least; (see, among others, Beck et al., 2004, 2010; Bogal-Allbritten, 2013; Bochnak, 2015).

[3]In this paper, the diacritics on sentences reflect my choices which, in some cases, differ from those of the authors whose work is under discussion. "?" is used throughout to indicate a felt anomaly related to meaning; when any other diacritics are used, flags are supplied to indicate how are they are meant to be interpreted.

[4]The difference between *tall* and *full* is often referred to as the relative/absolute distinction in gradable adjectives, and was apparently first noticed by Unger (1971). Not all of the relevant tests are reviewed here; see Rotstein and Winter (2004), Kennedy and McNally (2005), Lassiter (2011, 2017), and Klecha (2014) for tests involving proportional modification with *half, 90%*, and *mostly*, entailments between sentences with positive adjectives and their intuitive antonyms, and others. Of note in light of the squishiness of judgments in this domain is Kennedy's (2007) suggestion that *perfectly* (maximizer) and *slightly* (minimizer) are generally best at showing the relevant interpretive patterns across a broad array of gradable adjectives (Kennedy, 2007, p. 34).

[5]Of course, we assume that speakers are generally competent in their language and know what the expressions in their language are about (or, at least, linguists tend to assume this); compare, for example, Putnam (1975) and Chomsky (1995).

[6]Sassoon (2011) hypothesizes that *slightly* works differently—roughly, it picks out different standards depending on whether the property denoted by the adjective is stable (i.e., individual-level) or not (i.e., stage-level)—by noting that all relative adjectives seem to have an intuitive zero, including that associated with *(in)expensive*. Extended consideration of Sassoon's interpretation for *slightly* would complicate the narrow point I want to make in the text but could itself be used to raise the same general issue: how should we independently determine what counts as a stable vs. non-stable property, independently of linguistic tests like compatibility with adverbs like *frequently*, *rarely*, etc.? See also Toledo and Sassoon (2011) and Solt (2012).

(7) ? Mary is taller than Sam isn't.

However, the scale associated with *full* apparently does have a scalar maximum [recall (5a)], yet (8) is deviant (cf. Lassiter 2011, p. 12).

(8) ? This glass is fuller than that glass isn't.

Lassiter furthermore points to adjectives like *tall* that are intuitively lower-bounded yet fail to pass tests for minimal scalar points. For example, compatibility with *slightly* is meant to track this scalar property, and yet, if that is the right analysis (see footnote 6), neither of the phrases in (9) mean what they should mean. That is, certainly neither (9a) nor (9b) should give rise to any felt anomaly, all else being equal; and, it seems to the current author, (9a) should just mean that Ann is really, really short, and (9b) that the watch is really, really cheap.

(9) a. ? Ann is slightly tall.
b. ? The watch is slightly inexpensive.

These issues did not go unnoticed by Kennedy (2007), who writes (p. 34–5),

> …why do the scales used by particular adjectives have the structure they do? For example, naive intuition suggests that the COST scale should have a minimal value representing complete lack of cost, just as the DIRT scale has a minimal value representing complete lack of dirt. However, the unacceptability of *??slightly/??partially expensive* and *??perfectly/??completely/??absolutely inexpensive* (cf. *slightly/partially dirty* and *perfectly/completely/absolutely clean*) indicates that as far as the gradable adjective pair *expensive/inexpensive* is concerned, this is not the case: the scale used by these adjectives to represent measures of cost does not have a minimal element.... The structure of a scale is presumably determined mainly by the nature of the property that it is used to measure, but the different behavior of e.g., *expensive/inexpensive* vs. *dirty/clean* suggests that this aspect of linguistic representation may diverge from what naive intuitions suggest.

Here, Kennedy raises the possibility of a divergence between intuitive judgments regarding the properties that are "out there" and their linguistic representations, whatever those might turn out to be.

Semanticists differ in their degree of comfort with this state of affairs. Lassiter (2010, p. 205) appears be worried, since: given our "intuitive assumptions about the nature of the scales in question, [the sentences in (10)] should be equivalent," contrary to fact.

(10) a. This pizza is completely inexpensive.
b. This pizza is free.

Klecha (2012), in contrast, is not worried: in his discussion of the scale structure associated with the epistemic adjective *likely*, he writes that "ultimately the 'intuitive scale' associated with an adjective does not always align with its lexical scale" but "nor should we expect it to"; instead, we may acknowledge what "may seem like counterexamples," but, "just because these intuitive scales" have some apparent bounding property, "we should not conclude that the lexical scales" do too (p. 11)[7].

In general, the position expressed most stridently here by Klecha is common in linguistic semantics, but I have only found it discussed explicitly in the context of evaluating whether natural language semantics is best pursued as a theory that interfaces with metaphysics as opposed to something else. The predominant view arising in these discussions appears to be that instead of building a theory of how linguistic expressions compositionally relate to the (real) world, we build a theory of how linguistic expressions compositionally relate to the world as we talk about it. Therefore, we assume a world that is as language suggests it to be, not as it actually is, and the interfacing theory for semantics is "natural language metaphysics" rather than (real) metaphysics (Bach, 1986; Bach and Chao, 2012; cp. Moltmann, 2017). The problem with such a position, I contend, is that it amounts to a refusal to say what semantics properly interfaces with; under these conditions, its theoretical statements cannot be evaluated for truth and falsity. This renders semantics non-scientific.

More generally: if a semantic theory aims to explain certain semantic judgments in terms of something else—such as what those expressions are about—then it had better be that we have an independent theory of what expressions are or can be about. In other words, the theory has to respect both our semantic and metaphysical intuitions and provide for a way of resolving mismatches where they are found. Much caution is warranted. For present purposes, relevant modifiers and gradable adjectives might be polysemous[8], and in ways that are not entirely predictable; this requires antecedent caution in interpreting the results of our linguistic tests. And, even supposing that we can fix on the appropriate senses for the purposes of making judgments, not all of the tests work all the time, "for apparently idiosyncratic reasons" (Kennedy, 2007, p. 34).

I think there is a way to account for semantic anomaly and to respect our independent judgments of what our expressions are about. However, much of the hard work of showing how to do it has not yet been done. This paper will not do all of that work, but it aims to contribute to the bigger project by focusing in on explanation in this corner of degree semantics. Section 2 gives a number of additional examples of theoretical posits proffered within that framework and describes some of the

[7]Lassiter, for his part, supposes not that the tests fail to show what they purport to show but just that the relevant generalizations are weaker than their architects supposed: the inference from *completely A*, for adjective *A*, to a scalar endpoint holds, but *A*'s associating with a scalar endpoint does not guarantee the felicity of *completely A*. In other words, the generalization is a conditional one rather than a biconditional one. (Lassiter and Goodman, 2013 offer a very different approach to the relative/absolute distinction.) If so, this would not be so surprising, though it does pose its own explanatory challenges. The situation is analogous to that in the mass/count literature with respect to lexical specifications being overridden by, for instance, the semantic commitments of plural morphology (see e.g., Gillon, 1992, 2012).

[8]Lassiter notes some of the many senses that the maximizing modifier *completely* can take on, obscuring the results of those tests: it can function as a marker of "emphasis, correction, or high speaker confidence" (Lassiter, 2011, p. 13). Beltrama's (2018) study provides an interesting contrast between *completely* and *totally*, which both have maximizing uses, but the latter expresses subjective intensification with adjectives that *completely* sounds awful with; compare ?*completely* tall and *totally tall*.

explanations those posits are used in service of. Section 3 returns to the question of what we understand our semantic theory to be doing—whether relating expressions to the (real) world, to the world as we talk about it, or, instead, to other areas of cognition. And, section 4 takes a stab at a specific positive proposal. This proposal—interpretation in two steps—holds some additional appeal in that it provides some resources for capturing polysemy.

2. MATTERS FOR INTERPRETATION

This section briefly lays out the essentials of a degree-based compositional semantic theory, with special attention to the hypothetical nature and variety of things that linguistic expressions are about as they are recruited for semantic explanation[9]. These roughly fall into three categories that are not entirely independent: degrees, the scalar relations that order them, and the measure functions that relate entities to scales. I lay out some of the claims here but will not, for the most part, comment on their interrelations.

Beginning with degrees themselves, a first distinction found in the literature is between whether degrees should be understood to be primitive (i.e., not reducible to abstractions based on other objects; the default assumption) or as labels for equivalence classes of objects (as in Cresswell, 1976), possible objects (Schwarzschild, 2013), or of states (Anderson and Morzycki, 2015), etc. Appeal to degrees simpliciter, or to aspects of their nature, has important consequences for the data coverage of a degree-based theory. Additionally, while I will not discuss it here, their importance for linguistic theory is supported by the need for an account of movement-like properties in *than*-clauses, which receives a natural account in terms of abstraction over degrees; see Kennedy (2002) for extensive discussion and references.

With the introduction of degrees, we are able to explain certain basic data concerning the interpretation of comparatives with *-er*/*more*, *as*, etc. In a degree semantic setting, such comparative constructions are typically analyzed in terms of a greater-than relation between two degrees d and d', such that *x is A-er than y* is true only if x is mapped to a higher degree on the scale associated with A than y is. Some adjectives associate with the same scale, or so it is supposed based on consideration of what have come to be called "subcomparatives" like (4), repeated as (11) below.

(11) a. The ladder is taller than Ann's son is wide.
b. ? Ann's glass is fuller than the ladder is tall.

These examples show that while distinct adjectives A and A' occur in the matrix and *than*-clauses of the comparative, not everything goes: comparatives like *x is A-er than y is (A′)* are true just in the case where x is mapped to a higher degree on the scale that is common to A and A' than y is[10]. Since (11a) is perfectly acceptable and interpretable while (11b) is not, we may posit thereby that (11a) involves adjectives that share a scale of length whereas there is no common scale for (11b).

A second cut is in whether the comparative morphology relates degrees simpliciter (i.e., degrees as points, ordered by some $\leq$) or convex sets of such degrees (i.e., degree intervals, ordered by an inclusion relation $\sqsubseteq$). Interpreting comparatives as essentially relating scalar intervals helps to explain otherwise puzzling data relating to quantificational noun phrases in *than*-clauses (see especially Schwarzschild and Wilkinson, 2002; Fleisher, 2016). Assume that we have Ann and 10 other people, such that 5 people are shorter than Ann and 5 people are taller. Under these circumstances, (12) is intuitively false.

(12) Ann is taller than everybody else is.

Yet, supposing that the derivation of *the* degree named by a *than*-clause involves some calculation using a set like $\{d$: everybody but Ann has d-height$\}$, there is no way to predict this judgment correctly; this difficulty can be overcome by positing that the calculation involves certain sets of degrees (see Schwarzschild and Wilkinson, 2002 for details).

Kennedy (2001) builds on the idea that comparative constructions involve the manipulation of scalar intervals but extends it so that these may come in positive and negative varieties[11]. In particular, he aims to account for the fact that, even if two adjectives share a dimension, the comparative form is unnatural if the two adjectives are opposite in polarity; see, for instance (13).

(13) ? The ladder is longer than the doorway is short.

Kennedy explains the anomaly of examples like (13) by positing that *long* relates the ladder to a positive interval—one stretching from 0 length to the length of the ladder—while *short* relates the doorway to a negative interval—one stretching from the length of the doorway up to infinity. Since there is, in principle, no possible inclusion relation between such degrees, Kennedy suggests, the comparative form is anomalous[12,13].

More can and has been said about degrees per se, but present purposes concern what has been said of the scales that order them. The most lauded aspect of scalar structure in the degree semantics literature in recent years concerns whether the relevant scale has certain privileged elements—an upper bound or maximum and a lower bound or minimum (Rotstein and Winter, 2004; Kennedy and McNally, 2005). A battery of tests, some of which were cited in the previous section, are thought to

[9]For more detailed overviews of degree semantics, see Kennedy (2006), Schwarzschild (2008), Wellwood (2019), chapter 2.

[10]"Regular" comparatives like *Ann is taller than Bill is* represent the identity case, where $A = A' =$ *tall* (see Bresnan, 1973).

[11]In Kennedy's technical implementation, given some privileged point d, a positive degree interval is one that starts at the scalar minimum and extends up to d, while the (near-) complementary negative degree interval is one that begins at d and extends upwards to infinity. This implementation is at odds with some of the details of later developments in modeling scale structure; see below.

[12]This is a species of triviality argument: since, in virtue of its syntax-semantics correspondences, such a sentence will never evaluate to true or false, it is unacceptable; see Gajewski (2002) for extended discussion of this type of theoretical reasoning.

[13]Büring (2007) points out that the phenomenon of "cross-polar anomaly" is actually somewhat more restricted, noting that it only occurs if the negative adjective appears in the *than*-clause; contrast (13) with *The doorway is shorter than the ladder is long*, which is reported to be acceptable. Büring suggests not that the lack of anomaly is a counter-example to Kennedy's theory, but that it reveals syntactic decomposition of negative adjectives; see his paper for details.

diagnose whether an adjective or antonymic pair associate with different scales in a typology like that displayed in (14), where the examples given are instances of hypothesized antonymic pairs whose shared scale bears the relevant properties (from Kennedy and McNally, 2005).

(14) Hypothesized scalar typology
   a. Open (e.g., *tall*, *short*)
      No scalar minimum or maximum
   b. Lower-closed (e.g., *bent*, *straight*)
      Only a scalar minimum
   c. Upper-closed (e.g., *certain*, *uncertain*)
      Only a scalar maximum
   d. Totally closed (e.g., *full*, *empty*)
      Both a scalar minimum and maximum

One last arena in which degree semantics makes substantial demands of ontology or conception in its explanations concerns measure functions, which introduce a relation between measured entities and the scales that represent their measures[14]. Given basic assumptions of degree-semantic theories, we need not expect any particular correspondences between (call it) the structure of the entities measured and that of the scales used to measure them. And while it may appear that we do not see such a correspondence, in some cases we certainly do.

For one such case, consider the comparatives in (15).

(15) a. Ann had more mud/intelligence/heat than Bill did.
   b. Ann bought heavier/darker/tastier mud than Bill did.

With bare *more* in (15a)[15], the meaning of the noun determines dimensionality: *more mud* can be used to express a thought about relative volume or weight, but not about heaviness, darkness, or tastiness, unlike (15b). Meanwhile, *more intelligence* and *more heat* cannot, or so it seems, be used to express a thought about relative volume, weight, heaviness, darkness, or tastiness, etc; rather, their dimensions are specific to whatever *intelligence* and *heat* describe.

The facts are parallel in the verbal domain; consider (16).

(16) a. Ann ran/shone/sped up more than Bill did.
   b. Ann ran faster/more gracefully than Bill did.

To say that (15a) involves instances where there must be alignment between what is measured and how it is measured (i.e., what scale is used to represent the measurement) is to say that the dimensions for comparison with bare *more* uniformly appear to preserve certain formal features that the measured domains appear to have. That is, many authors have described the relevance of mereological or part-whole relations on the extensions of (at least) phrases like *mud* and *run*: whatever *mud* can be truthfully used to describe, it also can truthfully describe arbitrary subparts thereof (Cartwright, 1975; Link, 1983, and many others); and whatever *run* truthfully applies to, it also truthfully applies to arbitrary subparts thereof (Taylor, 1977; Bach, 1986, and many others). These patterns of application can be modeled by partial orders on portions of mud or stretches of running, and it is the strict ordering relations that are preserved in the mapping to degrees (Schwarzschild, 2002, 2006; Nakanishi, 2007): smaller portions of the mud have smaller volume or weight measures but not smaller temperature measures; smaller stretches of running activity measure less by duration or distance but not by speed.

This is not the only arena in which structure-preserving relations between distinct ontological or conceptual domains have been important in degree semantics (see Hay et al., 1999; Kennedy and McNally, 2005; Kennedy and Levin, 2008; Piñón, 2008). It has been supposed that there are non-trivial correspondences between the scale structure associated with a gradable adjective and the telicity profile of its corresponding deadjectival verb. Of particular interest for present purposes is the observation that telic verbal descriptions track scalar maxima associated with their adjectival core (if available) while atelic verbal descriptions track derived scalar minima (see Kennedy and Levin, 2008 for discussion and references).

Relevant data include 'degree achievements' (Dowty, 1979). Among the pertinent observations are: (i) some deadjectival verbs show variable telicity, and (ii) some are only atelic. With respect to (i), verbs such as *to cool* are said to be variably telic in that they are compatible both with telic (*in X time*) and atelic (*for X time*) modifiers. Interestingly for our purposes, depending on the modifier they show different implications: (17a) with a telic modifier suggests that the soup became maximally cool, while (17b) with an atelic modifier merely implies that the soup became cooler than it was before.

(17) a. The soup cooled in 10 min.
   b. The soup cooled for 10 min.

With respect to (ii), degree achievement verbs like *to widen* are only acceptable and interpretable with atelic modifiers, requiring only a minimal change in degree; compare (18a) and (18b).

(18) a. The gap between the boats widened for a few minutes.
   b. ? The gap between the boats widened in a few minutes.

In Kennedy and Levin's analysis (see also Kennedy, 2012; McNally, 2017), the truth of such predications depends on the positive interpretation of their adjectival core; their truth conditions, in turn, are derived via a mapping from the scalar

[14] Recent research has toyed with revising this basic assumption, analyzing gradable adjectives in terms of properties of states rather than in terms of degree functions (see Fults, 2006; Wellwood, 2012, 2015; Baglini, 2015; Pasternak, 2017; Cariani et al., 2018; Glass, 2019).

[15] I say "bare" because it does not appear with a lexical adjective or adverb. In (15a), there is a "nominal comparative," but the facts are parallel for verbal comparatives, as I show below. Such cases plausibly involve a functional quantificational element corresponding to (some occurrences of) English *much*, which plays the role of introducing measure functions; cf. Bresnan, 1973; Wellwood et al., 2012.

structure associated with the adjective into the event structure associated with its embedding verb phrase. In particular, variably telic predicates involve interpretation relative to scalar maxima (telic) or contextual standards (atelic). As in the positive adjectival form, whether the predication is maximal or not depends, by default, on whether the adjective's scale has a maximal element. Crucial for our purposes is the idea that the scale associated with the base adjective, call it S_A, is mapped onto a scale, call it S_{Δ_A} that measures degree of change. These derived scales, it is supposed, all have a minimal element (corresponding to the degree of the object along S_A, before the change occurs), but they only have a maximal element if S_A has a maximal element.

Thus, *to cool*, based on the upper-closed S_{cool} (witness the acceptability of *completely cool*), can be interpreted as telic—where an object x reaches the maximal degree of change possible along the relevant dimension, namely when x has reached the maximal degree of coolness—or atelic—where x reaches some change greater than the minimum, that is, where x was along S_A at the initiation of the change event. In contrast, *to widen* has only the atelic interpretation because the scale measuring change has exactly that kind of minimum—the degree to which x is wide at the start of the change event—but it fails to inherit a maximum from where it fails to exist in S_A.

What should be clear is that quite a lot of the theoretical description of what is going on in this corner of language involves assumptions about the sorts of things quantified (degrees), how they are ordered, and the presence or absence of "special elements" in those orderings (scales), in addition to structure-preserving relationships between the degrees used to represent measurement and the entities so measured. What I want to know is: apart from the evidence of semantic analysis itself, however copious that evidence, what independent tests are there for the adequacy of the attendant semantic explanations? Precisely to the extent that those explanations depend on independent features of ontology or conception, we require the details from an independent theory that describes those features.

3. THE MEANING RELATION

For concreteness, let us regard some of the statements formalizing the theoretical claims presented in the previous section. For instance, Kennedy and McNally (2005) derive the interpretation of an adjectival phrase consisting of a gradable adjective like *expensive* and the silent positive morpheme, POS (responsible for linking entities with a contextual standard for the target adjective; see discussion and references in their paper). In addition to its role in selecting a standard in c for the adjective, itself interpreted as in (19a), POS has the function of binding the degree argument introduced by that expression, (19b).

(19) a. $[\![\text{expensive}]\!] = \lambda d\lambda x.\mathbf{expensive}(x) = d$

b. $[\![\text{POS}]\!] = \lambda g\lambda x.\exists d[\mathbf{standard}(d)(g)(c) \wedge g(d)(x)]$

c. $[\![\text{POS}]\!]([\![\text{expensive}]\!]) = \lambda x.\exists d[\mathbf{standard}(d)([\![\text{expensive}]\!])(c) \wedge [\![\text{expensive}]\!](d)(x)]$

The result of this local computation is the property in (19c): it is a property true of individuals who measure some degree of expensiveness greater than the standard for expensiveness in c. The general schema in (20) highlights where and how the scale structure associated with the adjective might come into play: as the degree relation (e.g., the interpretation of the gradable adjective) acts as an argument to the **standard** function, which, for reasons described in Kennedy and McNally's paper, will default to the maximum when the degree relation has a maximum degree in its range, etc.

(20) $[\![\text{POS}]\!]([\![\text{A}_{\max}]\!]) = \lambda x.\exists d[\mathbf{standard}(d)([\![\text{A}_{\max}]\!]) \wedge [\![\text{A}_{\max}]\!](d)(x)]$

Kennedy and McNally, like the other authors whose work is considered in any detail here, assume a semantic framework like that laid out in Heim and Kratzer (1998), which is properly read as implying nothing more than a computational-level description of what speakers know when they can be said to know some piece of their language. Statements like those in (19) and (20) reflect a hypothesis about what such speakers know: they know the correspondences established by the interpretation function, $[\![\cdot]\!]$. The manner of specification for the terms "on the right" of equations involving $[\![\cdot]\!]$ usually are not intended to be taken as theoretically loaded qua symbols (see Dowty, 1979; Williams, 2015 for discussion of "semantic representations"). Nonetheless, if knowing one's language implies knowing such statements, and if such statements involve properties of things which are not obviously properly linguistic, then the theory depends for its explanations on the independent determination that those things in fact have those properties and that competent speakers know this[16].

How do we determine whether the relata "on the right" have the properties our theories need them to have? There are different ways one can approach this question and thus different ways one may begin to get beyond the explanatory impasse. The first bites the bullet and supposes that semantic theory interfaces with research in metaphysics—the study of what there is. The second is impervious to the bullet and supposes that semantics proceeds in isolation, describing and depending on properties of things needed for semantic analysis but without any attendant commitment to whether those things have any independent existence, whether "out there" (metaphysical) or "in the head" (cognitive). The third dodges the bullet by supposing that, despite our theoretical talk of establishing word-world relations, our theory is primarily geared toward describing a relation between expressions and elements of non-linguistic cognition.

[16]This interpretation dovetails with some of the few explicit statements of what such semantic theories are committed to: for example, Higginbotham (1985) writes, "Semantic theory proceeds from assumptions both about the nature of syntactic structures and about the nature of semantic values" (p. 553), and Bach (1986) writes, "I understand 'semantics' in the sense of a theory of the relationship between language and something that is not language" (p. 574).

I will suggest that pursuing the third option provides our best hope of overcoming some of the challenges posed by our case study.

3.1. Language and the World

Taking the semantic theory to be truth-conditional—i.e., as specifying, for each well-formed sentence *S* of the language, what it would take for *S* to be true, in this or some possible world—takes it to depend, in non-trivial ways, on what is or could be true (see Travis, 1996). What does this mean for present purposes? Given a sentence *S*—say, *Ann's glass is completely full*—the theory pairs *S*, via $[\![\cdot]\!]$, with a statement to the effect that *S* is true only if the scale associated with *full* has a maximal point, $d_{\max}$, and Ann's glass measures full to $d_{\max}$[17]. Among other things, this theory entails that there exists a scale of fullness that has certain properties. To some ears, this may sound straightforward and unimpeachable. However, if counter-examples like those discussed in section 1—those showing mismatches between our intuitions about which sentences are anomalous and what scales are like—are thoroughgoing and pervasive enough, a theory that depends on "what there is" for its evaluation may quickly come under threat.

The idea that "the meaning relation" establishes how expressions are related to the things we use our expressions to talk about has unobjectionable roots. First, as speakers, we use language to talk about the world, and the primary source of evidence for semantics comes from the correspondences (or lack thereof) between the way the world is and how we use our sentences to say that it is. Second, as theorists, we follow Lewis's and Cresswell's advice (by way of Partee, 1995): we broker the mystery of meaning by finding something that does what meanings do and study that; and, minimally, meanings make a difference in truth; so, we should be able to inform any study of meaning by way of the study of truth conditions.

A general problem is that specifying "the conditions under which *S* would be true" will involve specifying far more than what we want to attribute to the linguistic object, *S*, alone—and that is what a semantic theory aims to target. The trouble is easy enough to see in puzzle cases: considering the anomaly of a sentence like *The rock thinks it's raining*, Chierchia and McConnell-Ginet write, "the oddness seems linked more to the structure of the world than to facts about linguistic meaning: rocks just aren't the kind of thing that thinks, as it happens, but this seems less a matter of what *rock* and *think* mean than a matter of what rocks and thinking are like" (p. 48). But it is also plain in mundane cases: detailing the conditions under which *It was raining outside at noon on 10/4/2019* would be true would require, in fact, a catalog of how the whole world at a particular moment (and the history leading up to that moment) came to instantiate the state of affairs said to have been instantiated.

Returning to the central problem: how can we know what the properties of the things "on the right-hand side" are like such that we can evaluate our semantic theory for its own truth or falsity? Considering the theory to relate expressions to the world, we have two options for independent theories that might do the job of independently specifying worldly properties: physics or metaphysics. If a criterion for a semantics-as-science is that its interfacing theory is empirical, then we should go with physics. But this will not do; as Hobbs (1985) succinctly puts it (p. 20),

> Semantics is the attempted specification of the relation between language and the world. However, this requires a theory of the world. There is a spectrum of choices one can make in this regard. At one end of the spectrum—let's say the right end—one can adopt the "correct" theory of the world, the theory given by quantum mechanics and the other scientists. If one does this, semantics becomes impossible because it is no less than all of science, a fact that has led Fodor (1980) to express some despair. There's too much of a mismatch between the way we view the world and the way the world really is. At the left end, one can assume a theory of the world that is isomorphic to the way we talk about it. . . . In this case, semantics becomes very nearly trivial.

I do not think semantics is trivial. But how do we ensure that it is not? When semanticists are explicit about the question of an independent, interfacing theory, they tend to assert that we do not need one and that we can do just fine with a model of things we "talk *as if*" there are. But this will not do either, as I discuss next. A quite different alternative, of course, would involve reinterpreting the statements in our semantic theory as reflections of (or abstractions over) how our expressions relate to categories of mind; I discuss this in section 3.3.

3.2. "Talk *as if*"

Some contend that the entities posited in semantic explanations have an existence entirely within the theory and do not (and should not) depend for their properties on an independent theory[18]. Thus, semantics traffics in what we talk *as if* there is (Bach, 1986; Bach and Chao, 2012; cp. Moltmann, 2017) and understands that talk neither in metaphysical nor cognitive terms. This position has come to be called "natural language metaphysics" (NLM; Pelletier, 2011 calls it "semanticism"). This position does have points in its favor, as reviewed below. But none of these overcome its inherent scientific deficiency.

For Pelletier (2011), the main considerations in favor of "talk *as if*"/NLM have to do with apparently extensionally equivalent referents for terms that otherwise have been thought to be loaded with ontological commitment. For example, many truth-conditional approaches to the mass/count distinction suppose that it is characteristic of mass terms like *water* that they refer divisively, while count nouns like *cup* lack this sort of reference; as a reminder, for anything that *mud* applies to, *mud* also

[17] In this framework, semantic theory describes $[\![\cdot]\!] : E \rightarrow Z$, *E* the set of morphosyntactic objects, *Z* of worldly entities. This is so in the Montagovian tradition á la Heim and Kratzer (1998), where I draw $[\![\cdot]\!]$ from; a weaker, relational (but still truth-conditional) approach is taken in the Davidsonian tradition á la Larson and Segal (1995), where the relevant relation is called "Val"; see also Martin (1958), as well as works by Higginbotham, Boolos, Pietroski, and Schein (Schein p.c.).

[18] A *Frontiers* reviewer points out that that there may be language-internal reasons why certain (classes of) expressions behave in unexpected ways, for instance constraints on how lexicalization carves up conceptual space (see footnote 23). However, the possibility of recruiting such explanations will plainly depend on one's foundational assumptions, which leads us right back to the present matter of which foundations we should accept for the purposes of semantic theorizing.

applies to any of its arbitrary subparts (mass; divisive reference), but for anything that *a toy* applies to, *a toy* does not also apply to its arbitrary subparts (count; non-divisive reference). In mereological approaches to the mass/count distinction (e.g., Cartwright, 1975; Parsons, 1979; Link, 1983), these referential profiles are attributed to ontological differences between what we might intuitively think of as "substances" and "objects."

Pelletier (2011) takes issue with this because it just does not seem that a semantics based on wholes and subparts ad infinitum for the mass noun *water* jives with what we know about the stuff, water. After all, water has smallest parts—H_2O molecules. He writes (p. 26)[19].

> A standard defense of the divisiveness condition in the face of these facts is to distinguish between "empirical facts" and "facts of language." It is an empirical fact that water has smallest parts, it is said, but English does not recognize this in its semantics: the word *water* presupposes infinite divisibility.
> It is not clear that this is true, but if it is, the viewpoint suggests interesting questions about the notion of semantics. If *water* is divisive but water isn't, then water can't be the semantic value of *water* (can it?). In turn this suggests a notion of semantics that is divorced from "the world", and so semantics would not be a theory of the relation between language and the world. But it also would seem not to be a relation between language and what a speaker's mental understanding is, since pretty much everyone nowadays believes that water has smallest parts. Thus, the mental construct that in some way corresponds to the word *water* can't be the meaning of *water* either.

I will address the specific concern about there being a unique construct that *water* associates with in section 4. But Pelletier cites other empirical considerations that, he contends, militate semantic theory toward agnosticism: for one, in English and in other languages there are pairs of words that are drawn from the mass and count sides of the distinction and yet "the items in the world that they describe seem to have no obvious difference that would account for this" (p. 26), like *spaghetti* and *noodles*. And do we really think about what is on the plate differently depending on the word we choose? For another, citing data from Chierchia (1998), Pelletier notes that while both English and Italian have both mass and count noun forms corresponding to *hair/s*, in English you say *I cut my hair* but in Italian you say (the equivalent of) *I cut my hairs*; yet, clearly "It would seem that the same activity is described no matter where the barber is doing the work" (p. 29).

As an aside, I think there are reasons to suppose that these problems in particular do not loom as large as might seem, particularly if one posits a derivational—rather than merely lexical—account of the distinction between mass and count occurrences of nouns (cf. Borer, 2005). If *mud*, for example, amounts to meaning "stuff that we call mud" and *muds* amounts to "a plurality of entities, each of which is constituted by some stuff that we call mud," do some of these worries evaporate? Regardless, it is unsatisfying in the extreme to conclude that we should thereby land firmly on the side of NLM, where nothing worldly nor conceptual should be recruited in order to help explain the mass/count distinction. After all, there certainly are robust correlations between the grammatical mass/count distinction and the notional object/substance distinction that will need explaining (cf. Rips and Hespos, 2015).

NLM amounts to a refusal to say what the interfacing theory with semantics is or should be. It thus puts semanticists in an uncomfortable place: assuming, as most do, that our compositional theories are bounded "on the left" by syntactic and morphological theory, we nonetheless resist bounding our theory "on the right" by anything at all. If there was nothing else that could be said, so be it. But it cannot be that we avoid committing simply in order to avoid making bad predictions.

3.3. Language and the Mind

What is left? What remains is the view that the study of language begins with its study as a faculty of mind and characterizes the knowledge recruited by that faculty during linguistic understanding and production. Semantic theory bridges the language faculty with other faculties of mind. On such a view, semantics tracks morphosyntactic structure "on the left" and non-linguistic cognition or conceptual structure on the right[20]. (Then, if we are lucky, the concepts and categories "on the right" can be, in their turn, related to aspects of real reality.)

This view requires, of course, that the theory take the form of a relation between two levels of structured representation. Computation in any form is syntactic, and the nature and structure of the symbols computed over play an important role in what computations can be performed[21]. When Lewis (1970) famously dismisses early attempts to characterize semantic theory as a relation between two languages—say, English and "Markerese" (e.g., Katz and Fodor, 1963)—he does so because of a strong prior commitment that semantics *as such* implies a relationship between language and the world. Referring to the structured language-like outputs of the 'projection rules' in a generative semantic model like Katz and Fodor's as "semantic markers," Lewis writes (p. 18),

> Semantic markers are symbols: items in the vocabulary of an artificial language we may call Semantic Markerese. Semantic interpretation by means of them amounts merely to a translation algorithm from the object language to the auxiliary language Markerese. But we can know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely, the conditions under which it would be true. Semantics with no treatment of truth conditions is not semantics.

[19] This worry applies if we interpret the divisiveness condition very strongly, such that any divisive N, if it applies to some stuff, also applies to any arbitrary subpart of that stuff. See Bunt (1979) and Champollion (2017), among others, for important discussion.

[20] See Partee (2018) for discussion of the history of tension here; she cites Jackendoff (1983) and Fodor (1975) as early exemplars of the current position.

[21] For an accessible introduction to basic concepts in the computational theory of mind and the consequences of a representational format in particular (see Gallistel and King, 2010).

However, he does point to a way in which he may be okay with such theories; so long as they make a provision for *real* semantics, continuing,

> Translation into Markerese is at best a substitute for real semantics, relying either on our tacit competence … as speakers of Markerese or on our ability to do real semantics at least for … Markerese.

Remarkably, Chomsky (1989) appears to think this is precisely how it goes—that the phenomena a Lewisian semanticist is characterizing is a step removed from language proper, writing (p. 324),

> [the first] step in the process of interpretation …should be considered to be in effect an extension of syntax, the construction of another level of mental representation beyond LF ['Logical Form'], a level at which arguments at LF are paired with entities of mental representation, this further level then entering into 'real semantic interpretation.'

Allowing for such a "two-step" interpretation would allow the theorist to be an internalist about linguistic meaning but an externalist about semantics, if that term is reserved for theories of how expressions (in whatever language) relate to the world.

Pelletier (2011) expresses concern that shifting the work of semantic theory wholly 'inside the head' would take us too far away from the data on which the theory is based, namely, communication: "For one thing, it is difficult to see how mutual understanding can ever be guaranteed or even achieved with such a view" (p. 33). Worse, "it is hard to see how any truth-conditional account could be involved in conjunction with internalism" about meaning (Pelletier, 2011). Yet, while at least Jackendoff (1984, 1994, 2002) has attempted to show how we might model the first step, only recently have there been stirrings from within the truth-conditional camp that would support two-step interpretation at all. As examples, though, Glanzberg (2014) inches in this direction, supposing that the primary data result from "features of meaning represented within the language faculty, and features of extra-linguistic concepts." Pietroski (2010) takes things quite a bit further, as noted in some more detail below.

How could this help? If we take the intuitions of semantic anomaly like those in section 1 to indicate something about the relationship between morphosyntactic objects and elements of non-linguistic conceptualization, then it becomes an empirical matter what "scales" amount to—this cannot be stipulated in advance, and it need not track our folk intuitions about what such scales amount to "in the world"[22]. Our metaphysical judgments, just like our metalinguistic judgments, are the subject matters of different disciplines, interlocked in the explanation of how language is understood. More concretely, it will support a view in which the asymmetry between *completely full* and ?*completely tall* is explainable in terms of the nature and structure of the relation between language and conception, while our introspective intuitions about the nature of the associated scales are not.

I will next provide a sketch of how this might look from the perspective of formal semantics. Of primary importance, though, is that a view in which semantics primarily traffics in describing a language-mind connection invites cognitive psychology as a bound "on its right[23]." With it, we have an independent empirical theory that can restrict the nature and variety of the claims that semanticists can make with respect to what there is[24].

4. POSITIVE PROPOSAL

The approach I urge grounds at least some of our judgments of semantic anomaly in the relation between linguistic and non-linguistic cognition, but it grounds our judgments of truth and falsity in the relation between non-linguistic cognition and the world.

Where the traditional model in truth-conditional semantics (section 3.1) supports a boxology like that in **Figure 1**, I propose the finer-grained view in **Figure 2**. Assuming, not without controversy (see footnote 17), that the lines indicate functional relationships between one domain and another, the suggested picture characterizes semantic theories couched in $[\![\cdot]\!]$ terms as the composition of two functions, here *m* and *i* to evoke "meaning" and "interpretation," respectively. If all goes well, *i* will do what a truth-conditional semanticist wants $[\![\cdot]\!]$ to do, but it will assign truth conditions to Thoughts—structured representations internal to the mind that an animal may have quite independently of natural language (cf. Pietroski, 2010)[25]. In contrast, *m* will reveal, at least, the logical properties of natural language expressions and the classes of concepts relevant for their interpretation by *i*. I intend to locate anomalies like those discussed in section 1 at *m*.

Determining whether any given meaning-related phenomenon belongs in one or the other category is not

[22]For a recent study of early links between scalar language and independent cognitive systems for representing magnitudes (see Odic, 2018).

[23]Of course, the assumption that language relates to other faculties of mind is explicit in cognitive semantic approaches. Gärdenfors (2014) (and in many of his antecedent works) argues that lexicalization patterns are linked inextricably with the regions or bundles of regions in conceptual space. Assuming a sufficient independent theory of conceptual space, such a theory will make broad predictions about the sorts of meanings we expect to see lexicalized in human languages. Partee has long maintained that inattention to the lexicon in formal semantics is due to the fact that the problems of compositional or structural aspects of meaning are more tractable (see her 1980; 2018, for example).

[24]The following thought occurs to me, though it was not likely offered with an internalist conception in mind: "If we were to think of logic as relating to the structure of thought and of metaphysics as relating to the structure of reality, then logic would provide us with the most general traits of thought and metaphysics with the most general traits of reality" (Fine, 2012, p. 18). The semanticist's data concerning truth/falsity and entailment patterns seem to reveal, indeed, that natural language has a certain logic; I suggest that anomaly like the kinds of cases considered in this paper reveals (at least) our intuitive metaphysics.

[25]Pietroski argues that a number of features of language design can be explained in terms of a contrast between the format of linguistic meanings and extralinguistic Thoughts. A central example is that of predicate adicity: while much research in formal semantics suggests a monadic, eventive interpretation of verbs like *give* (i.e., $\lambda e.\mathbf{give}(e)$), there is evidence to suggest that the Thought associated with *give* sentences should nonetheless be analyzed with a triadic structure [involving, e.g., $\textsc{give}(x, y, z)$]; see his 2010; 2018. See also Gordon (2003), who provides evidence for the adicity of such event concepts in prelinguistic infants.

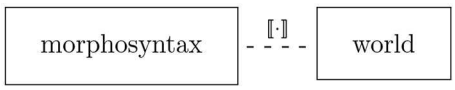


FIGURE 1 | The traditional model in truth-conditional semantics.

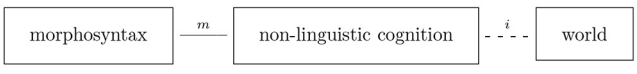


FIGURE 2 | A finer-grained model.

easy. However, it is possible that deeper probing of the nature of different judgments of (un)acceptability and (in)felicity could help. To begin thinking about this, we may first consider the well-known examples in (21) (Chomsky 1957).

(21) a. Colorless green ideas sleep furiously.
b. *Furiously sleep ideas green colorless.

Importantly, (21a) is a well-formed and acceptable sentence of English despite the impossibility of the state of affairs it describes, and the judgment that it is contradictory is almost beside the point. (21b), in contrast, is an ill-formed and unacceptable string of words in English—a thing for which the question of truth or falsity does not arise. Contrasting our target cases, (22a) like (21a) gives us no felt sense of anomaly, yet in my judgment (22a) presents as clearly and distinctly contradictory.

(22) a. The empty box is completely full.
b. ? The ladder is completely tall.

Unlike any of (21a), (21b), or (22a), (22b) is clearly unnatural and unacceptable, but this is apparently not due to any syntactic defect. Importantly, though, the question of truth or falsity does not arise for (22b) just as it does not for (21b).

What is needed, I submit, is a way of thinking about issues with the instructions for concept composition at play in (22b) but not in (21a). Within the general framework I advocate, at least two things must go right at *m* prior to evaluation of truth and falsity at *i*: (i) the sentence must be well-formed according to (at least) the morphosyntactic rules of the language, and (ii) the associated non-linguistic representations or concepts must themselves be well-formed[26]. (21a) and (22a) will, or so I shall suppose, meet both (i) and (ii)[27]. (21a) will run afoul of (i), and, I suggest, (22b) runs afoul of (ii). Such an explanation will require not only the familiar attention to (i) but serious acknowledgment of where the answers to (ii) may be found. How might we get there?

First, we may for simplicity's sake suppose that part of the meaning of lexical items is a "pointer" from within the language system to outside of it (Glanzberg, 2014). Then we can say that what determines whether a lexical item associates with a bounded scale (whether upper or lower) depends on what that lexical item points to and what relations and operations are defined for such concepts. (A central tenet of "core knowledge" approaches in psychology supposes that such knowledge comes in largely domain-specific packages, both in terms of representations and rules; see below.) If *tall* and *wide*, for example, point to a class for which length measures are defined, while *full* points to a class for which such measures are not defined, the explanation for the asymmetry in (4), repeated as in (23), can be explained in terms of these independent posits: (23b) invites the construction of a complex concept that cannot be evaluated for truth or falsity.

(23) a. The ladder is taller than Ann's son is wide.
b. ? Ann's glass is fuller than the ladder is tall.

Second, we must take quite seriously the types of restrictions that semanticists like lay down for the compositional requirements of expressions like *completely*, but understand them in a different way than previously. I suggest that we understand these requirements in terms of restrictions on the composition of concepts. More concretely, Kennedy and McNally (2005) suppose that (24) is a reasonable approximation of the semantic contribution of this modifier.

(24) $[\![\text{completely}]\!] = \lambda g \lambda x . \exists d [d = \mathbf{max}(S_g) \wedge G(d)(x)]$

As those authors write, "Assuming that the **max** function returns a value only for scales with maximal values, this modifier can combine only with gradable adjectives that have scales that are closed on the upper end" (p. 369). In the present framework, we may understand these specifications as restricting the space of concepts that *completely* can compose with. For a complete theory, we will want to know, of course, how to distinguish the concepts that are so composable from those that are not—and for this, we must turn to cognitive psychology.

To test our theories, we must take a hard and independent look at the neighboring cognition, as this is where empirical evidence for the nature and compositional structure of concepts can be sought. An easy place to start, I submit, is the cognitive and developmental psychology literature on core knowledge (for example, Spelke, 1998, 2003; Carey, 2009, and many others; see Strickland, 2016 for a related view)[28]. We know from this literature that, from the earliest stages of the development of humans as well as that of many other species, there exist domain-specific faculties of mind that undergird our intuitive understanding of what there is and how things work across a host of contentful categories: objects, events,

[26] Alternatively: the "instructions" for constructing those representations or concepts must be evaluable.

[27] If the difference in salience regarding their contradictoriness is real, one possibility is that this is due to the compositional distance between the pieces that compose to deliver the contradiction. In (21a), this point arrives as soon as *colorless* and *green* come together, for example, whereas in (22a), it arrives only once the subject is composed with the predicate.

[28] Partee (2018) seems to have a similar sort of investigation in mind, writing "…if we follow Burge (2010) in drawing insight from how perception works and how it gives (fallible) veridical knowledge prior to any "reasoning," we can see semantics, including at least parts of the difficult area of lexical semantics, as a particularly important and fruitful branch of psychology," suggesting further that "philosophy of language *need[s]* philosophy of mind for a resolution of some apparent problems in the foundations of semantics" (p. 190; her emphasis).

time, causation, agency, and more. The knowledge that partly constitutes each of these faculties is both highly specific and uniform across the species[29], and it is reasonable to suppose that the initial conceptual repertoire it provides restricts the available concept composition operations and scaffolds all later concept acquisition.

If we understand the formal statements of our semantic theory as encoding, in part, hypotheses about the kinds of representations and structures available in extralinguistic cognition, then we can test its predictions against what we know independently about extralinguistic cognition. In some cases, this can mean leveraging formal semantics as a source of hypotheses about representation. If the thematic or participant structure of events is important for a semantic theory, we can probe the nature and structure of our nonlinguistic event concepts in nonlinguistic tasks (e.g., Wellwood et al., 2015). If our theories require a privileged difference between object and substance predications, we can leverage the cognitive object/substance distinction (e.g., in the evaluation of *more* NP, see Barner and Snedeker, 2004; Odic et al., 2018). Where our theories say that the formal structure of objects and events is parallel, we can find ways of evaluating the psychological plausibility of the parallelism independently of language (e.g., Wellwood et al., 2018a).

One such arena of particular relevance for degree semantics is the literature on magnitude estimation, in which the Approximate Number System (ANS) is the most lauded[30]. The ANS is an evolutionarily ancient system that generates percepts of "numerosity," demonstrably in place in human children within the earliest time window in which it is possible to test (see especially Dehaene, 1997; Feigenson et al., 2004). Now, while ANS representations are ordered Gaussian distributions, which look different on the face from the set of discrete, ordered points required for cardinality comparisons in natural language, these two "scales" are isomorphic (e.g., Gallistel and Gelman, 1992; cf. Odic et al., 2015). And indeed, there is ample evidence that while the careful evaluation of a sentence like *Most of the dots are blue* tracks precise cardinality, speeded evaluation shows signs of the ANS (within and across individuals, across development, and across languages; see e.g., Halberda et al., 2008; Hackl, 2009; Pietroski et al., 2009; Lidz et al., 2011; Tomaszewicz, 2011).

In this light, we may consider how to address crosslinguistic differences like those noted in section 3, e.g., the apparent coextensivity of *spaghetti* and *noodles* despite their hypothetically distinct commitments to stuff vs. plurality. When English speakers use *spaghetti* as a mass term and Italians use it as a plural term, are they really thinking about what is on the plate differently?[31] This is an empirical question that can be tested. For example, if plural predications must be evaluated by number with *more* (see Wellwood, 2018 and references therein) while mass predications can but need not (see Barner and Snedeker, 2005 for experimental evidence), then we might expect *more spaghetti* to show more flexibility in its evaluation in English than in Italian when (say) number and volume are available as orthogonal options. Yet, we might appreciate a common *perception* by investigating how speakers view the images *independently of language* by constructing a comparable task that renders linguistic encoding unusable, e.g., by comparing similarity judgments of the same pairs of images, delivered while performing verbal shadowing[32].

On this general approach, the mismatches between our semantic and metaphysical intuitions pointed out in section 1 can be accommodated; since we distinguish the relations *i* and *m*, we may find restrictions in place at *m* (tracking our semantic intuitions) that are determined independently, and perhaps antecedently, to whatever we know at *i* (tracking our metaphysical intuitions). Recall, for example, the issue that our intuitive sense of the scale of cost—hypothetically that which is associated with adjectives like *expensive*—has a minimum element but *slightly expensive* does not mean what it should if the modification theory is correct. Our intuitions about what would count as a minimal element track *i*, but the anomaly we detect occurs already at *m*.

Importantly, a model of interpretation in two steps supports an account of polysemy[33], in which a single pointer (at *m*) involves a choice of resolution for the concept ultimately "fetched" (hence determining the input to *i*; e.g., Pietroski, 2018). Pelletier's (2011) worry about "the" semantic value of *water* could thus evaporate: we may have some early, core concept that we associate with the word but a different one after we do some science. Our early conceptual repertoire plays an important role in our cognitive economy throughout our lives and is likely responsible for endowing us with a naive concept of water that meets the divisiveness condition (cf. Prasada et al., 2002; Wellwood et al., 2018b). However, this repertoire does not restrict us from acquiring new concepts—e.g., one that is identical in extension with that of H_20—even if the two may ultimately be in conflict, metaphysically. This added degree of flexibility can similarly provide an angle on some of the cases discussed in the first half of the paper: perhaps language is wired by default to the sorts of concepts given to us biology—itself a matter of empirical discovery—that can differ from those that we acquire through reflection or experience.

Thus, our semantic intuitions might track properties of conception that are below what is available to introspection, while our metaphysical intuitions reflect a composite of (and sometimes tension between) our naive concepts and our more reflective ones.

[29] Indeed, they are likely responsible for our apparently species-level construction of a common experiential world (e.g., Jackendoff, 1994; Hoffman, 2009).

[30] Connections between the grammar of comparatives and the cognitive resources of magnitude estimation and comparison was suggested in quite another context in Fox and Hackl (2006).

[31] Incidentally, noteworthy Italian speaker Paolo Santorio, p.c., answers this question with a resounding "yes!".

[32] See Wellwood et al. (2016) and Wellwood et al. (2018b) for the use of a similarity-judgment task to get at the salience of numerical differences and Spelke (2003) for important results gleaned from verbal shadowing tasks.

[33] See, in this connection, Pietroski (2010) and Vicente (2012) for recent discussion, novel approaches, and citations to relevant literature.

5. CONCLUSION

I considered a case study in degree semantics and scale structure, leveraging putative counterexamples in this arena to advocate for a finer-grained model of semantic interpretation than is traditionally supposed within truth-conditional frameworks. Specifically, I offered the view that we can make sense of these counterexamples by assuming a model that divides interpretation into (at least) two steps. Semanticists are not in the business of formulating statements about how expressions compositionally relate to entities in the world but about how they compositionally relate to representations and operations in non-linguistic cognition. The outputs of the first step of interpretation—*m*—may themselves be submitted to truth-conditional evaluations that depend on what the world is like.

Semantic theory cannot only attend to what we talk as if there is, on pain of being rendered non-scientific. Instead, the two-step program integrates semantics within a tapestry consisting of necessary interdisciplinary links, wherein not only morphosyntactic theory but theories of conceptual structure inform theories of meaning and vice versa. As a bonus, the two-step program offers the kind of latitude that can support matters of polysemy, which will minimally be required for a complete account of the meaning of modifiers like *completely* (in addition to their guise as maximizers, they function as markers of confidence, etc.) More importantly, the possibility of a given lexical item pairing with more than one concept can help explain mismatches between our semantic and metaphysical intuitions.

The resulting view positions semantic theory as having a crucial role in furthering our understanding of the ways that the mind structures its experience. Semanticists theorize about all kinds of things that expressions may be "about"—in addition to objects, substances, and times, we posit events, processes, states, negative events, possible worlds, impossible worlds, and so on. Much of this talk would be news to psychologists, though there are already good case studies illustrating the payoffs for cognitive psychology of testing semantic posits as hypotheses about representation (for a very recent example, see Wellwood et al., 2018b). The approach I advocate thus invites semanticists to explicitly characterize their theory in such a way that it may be tested by these neighboring fields, and it invites psychologists to read our theories this way even when not so-intended. In this way, semantic theory can finally vindicate the idea that language is "a window into the mind."

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This research was supported by NSF BCS-1829225, awarded to AW.

REFERENCES

Anderson, C., and Morzycki, M. (2015). Degrees as kinds. *Nat. Lang. Linguist. Theor.* 33, 791–828. doi: 10.1007/s11049-015-9290-z

Bach, E. (1986). "Natural language metaphysics," in *Logic, Methodology and Philosophy of Science VII*, eds R. B. Marcus, G. J. W. Dorn, and P. Weingartner (Amsterdam: Elsevier Science), 573–595.

Bach, E., and Chao, W. (2012). "The metaphysics of natural language(s)," in *Handbook of Philosophy of Science, Vol. 14 of Philosophy of Linguistics*, eds R. Kempson, T. Fernando, and N. Asher (London: Elsevier), 175–196.

Baglini, R. (2015). *Stative predication and semantic ontology: A cross-linguistic study* (Ph.D. thesis). University of Chicago, Chicago, IL, United States.

Barner, D., and Snedeker, J. (2004). "Mapping individuation to mass-count syntax in language acquisition," in *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Societyw*, eds K. Forbus, D. Gentner, and T. Regier (Chicago, IL), 79–84.

Barner, D., and Snedeker, J. (2005). Quantity judgments and individuation: evidence that mass nouns count. *Cognition* 97, 41–66. doi: 10.1016/j.cognition.2004.06.009

Beck, S., Krasikova, S., Fleischer, D., Gergel, R., Hofstetter, S., Savelsberg, C., et al. (2010). "Crosslinguistic variation in comparison constructions," in *Linguistic Variation Yearbook 2009*, ed J. van Craenenbroeck (Amsterdam: John Benjamins Publishing Company), 1–66.

Beck, S., Oda, T., and Sugisaki, K. (2004). Parametric variation in the semantics of comparison: Japanese and English. *J. East Asian Linguist.* 13, 289–344. doi: 10.1007/s10831-004-1289-0

Beltrama, A. (2018). *Totally* between discourse and subjectivity: exploring the pragmatic side of intensification. *J. Semant.* 35, 219–261. doi: 10.1093/semant/ffx021

Bochnak, M. R. (2015). The Degree Semantics Parameter and cross-linguistic variation. *Semant. Pragmat.* 8, 1–48. doi: 10.3765/sp.8.6

Bogal-Allbritten, E. (2013). Decomposing notions of adjectival transitivity in Navajo. *Nat. Lang. Semant.* 21, 277–314. doi: 10.1007/s11050-012-9093-2

Borer, H. (2005). *Structuring Sense Volume I: In Name Only, Vol. 1*. Oxford: Oxford University Press.

Bresnan, J. (1973). Syntax of the comparative clause construction in English. *Linguist. Inq.* 4, 275–343.

Bunt, H. C. (1979). "Ensembles and the formal semantic properties of mass terms," in *Mass Terms: Some Philosophical Problems*, ed F. J. Pelletier (Dordrecht: Reidel), 249–277

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Büring, D. (2007). "Cross-polar nomalies," in *Proceedings of Semantics and Linguistic Theory 17*, eds T. Friedman and M. Gibson (Ithaca, NY: Cornell University), 37–52.

Burnett, H. (2016). *Gradability in Natural Language: Logical and Grammatical Foundations*. Oxford, UK: Oxford University Press.

Carey, S. (2009). *The Origin of Concepts*. Oxford studies in cognitive development. Oxford: Oxford University Press.

Cariani, F., Santorio, P., and Wellwood, A. (2018). *Confidence Reports.* Northwestern University m.s.

Cartwright, H. (1975). Amounts and measures of amount. *Noûs* 9, 143–164.

Champollion, L. (2017). *Parts of a Whole: Distributivity as a Bridge Between Aspect and Measurement*. Oxford studies in theoretical linguistics. Oxford: Oxford University Press.

Chierchia, G. (1998). Reference to kinds across languages. *Nat. Lang. Semant.* 6, 339–405.

Chomsky, N. (1957). *Syntactic Structures*. Berlin: Mouton.

Chomsky, N. (1989). *Lectures on Government and Binding*. Dordrecht: Foris.

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

Cresswell, M. J. (1976). "The semantics of degree," in *Montague Grammar*, ed B. H. Partee (New York, NY: Academic Press), 261–292.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. New York, NY: Oxford University Press.

Dowty, D. R. (1979). *Word Meaning and Montague Grammar, Vol. 7*. Dordrecht: Kluwer Academic Publishers.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends Cogn. Sci.* 8, 307–314. doi: 10.1016/j.tics.2004.05.002

Fine, K. (2012). "What is metaphysics?," in *Contemporary Aristotelian Metaphysics*, ed T. E. Tahko (Cambridge: Cambridge University Press), 8–25.

Fleisher, N. (2016). Comparing theories of quantifiers in *than*-clauses: lessons from downward-entailing differentials. *Semant. Pragmat.* 9, 1–23. doi: 10.3765/sp.9.4

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behav. Brain Sci.* 3, 63–73. doi: 10.1017/S0140525X00001771

Fox, D., and Hackl, M. (2006). The universal density of measurement. *Linguist. Philos.* 29, 537–586. doi: 10.1007/s10988-006-9004-4

Fults, S. (2006). *The structure of comparison: an investigation of gradable adjectives* (Ph.D. thesis). University of Maryland, College Park, MD, United States.

Gajewski, J. (2002). *L-analycity in Natural Language*. MIT (Unpublished).

Gallistel, C. R., and Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition* 44, 43–74. doi: 10.1016/0010-0277(92)90050-r

Gallistel, C. R., and King, A. P. (2010). *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. New York, NY: Wiley-Blackwell.

Gärdenfors, P. (2014). A semantic theory of word classes. *Croat. J. Philos.* XIV, 179–194.

Gillon, B. (1992). Towards a common semantics for English count and mass nouns. *Linguist. Philos.* 15, 597–639. doi: 10.1007/BF00628112

Gillon, B. S. (2012). Mass terms. *Philos. Compass* 7, 712–730. doi: 10.1111/j.1747-9991.2012.00514.x

Glanzberg, M. (2014). "Explanation and partiality in semantic theory," in *Metasemantics: New Essays on the Foundations of Meaning*, eds A. Burgess and B. Sherman (Oxford, UK: Oxford University Press), 259–292.

Glass, L. (2019). Adjectives relate individuals to states: evidence from the two readings of Determiner + Adjective. *Glossa* 4:24. doi: 10.5334/gjgl.552

Gordon, P. (2003). "The origin of argument structure in infants' event representations," in *Proceedings of the Boston University Conference on Language Development, Vol. 28* (Somerville, MA), 189–198.

Hackl, M. (2009). On the grammar and processing of proportional quantifers: *most* versus *more than half*. *Nat. Lang. Semant.* 17, 63–98. doi: 10.1007/s11050-008-9039-x

Halberda, J., Taing, L., and Lidz, J. (2008). The development of "most" comprehension and its potential dependence on counting ability in preschoolers. *Lang. Learn. Dev.* 4, 99–121. doi: 10.1080/15475440801922099

Hay, J., Kennedy, C., and Levin, B. (1999). "Scale structure underlies telicity in 'degree achievements'," in *Proceedings of Semantics and Linguistic Theory 9*, eds T. Matthes and D. Strolovitch (Ithaca, NY: CLC Publications), 127–144.

Heim, I., and Kratzer, A. (1998). *Semantics in Generative Grammar*. Malden, MA: Blackwell.

Higginbotham, J. (1985). On semantics. *Linguist. Inq.* 16, 547–594.

Hobbs, J. R. (1985). "Ontological promiscuity," in *Association for Computational Linguistics 23* (Chicago, IL), 61–69.

Hoffman, D. D. (2009). "The interface theory of perception: natural selection drives true perception to swift extinction," in *Object Categorization: Computer and Human Vision Perspectives*, eds S. J. Dickinson, A. Leonardis, B. Schiele, and M. J. Tarr (Cambridge, UK: Cambridge University Press), 148–166.

Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jackendoff, R. (1994). *Patterns in the Mind: Language and Human Nature*. New York, NY: Basic Books.

Jackendoff, R. (2002). *Foundations of Language*. Oxford: Oxford University Press.

Jackendoff, R. S. (1984). "Sense and reference in a psychologically-based semantics," in *Talking Minds: The Study of Language in Cognitive Science*, eds L. A. Bever, J. M. Carroll, and T. G. Miller (Cambridge, MA: MIT Press), 49–72.

Katz, J. J., and Fodor, J. A. (1963). The structure of a semantic theory. *Language* 39, 170–210. doi: 10.2307/411200

Kennedy, C. (2001). Polar opposition and the ontology of 'degrees'. *Linguist. Philos.* 24, 33–70. doi: 10.1023/A:1005668525906

Kennedy, C. (2002). Comparative deletion and optimality in syntax. *Nat. Lang. Linguist. Theor.* 20, 553–621. doi: 10.1023/A:1015889823361

Kennedy, C. (2006). "Comparatives, semantics of," in *Encyclopedia of Language and Linguistics*, ed K. Allen (Oxford, UK: Elsevier), 690–694.

Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguist. Philos.* 30, 1–45.

Kennedy, C. (2012). "The composition of incremental change," in *Telicity, Change, State: A Cross-Categorical View of Event Structure*, eds V. Demonte and L. McNally (Oxford: Oxford University Press), 103–121.

Kennedy, C., and Levin, B. (2008). "Measure of change: the adjectival core of degree achievements," in *Adjectives and Adverbs: Syntax, Semantics and Discourse*, eds L. McNally and C. Kennedy (Oxford: Oxford University Press), 156–182.

Kennedy, C., and McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81, 345–381. doi: 10.1353/lan.2005.0071

Klecha, P. (2012). "Positive and conditional semantics for gradable modals," in *Sinn und Bedeutung 16, Vol. 2*, eds A. Chernilovskaya, A. Aguilar Guevara, and R. Nouwen (Cambridge, MA: MIT Press), 363–376.

Klecha, P. (2014). *Bridging the divide: scalarity and modality* (Ph.D. thesis). University of Chicago, Chicago, IL, United States.

Larson, R. K., and Segal, G. M. (1995). *Knowledge of Meaning: An Introduction to Semantic Theory*. Cambridge, MA: MIT Press.

Lassiter, D. (2010). "Gradable epistemic modals, probability, and scale structure," in *Semantics and Linguistic Theory 20*, eds N. Li and D. Lutz (Ithaca, NY: CLC Publications), 197–215.

Lassiter, D. (2011). *Measurement and modality: the scalar basis of modal semantics* (Ph.D. thesis). New York, NY: New York University.

Lassiter, D. (2017). *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford, UK: Oxford University Press.

Lassiter, D., and Goodman, N. D. (2013). "Context, scale structure, and statistics in the interpretation of positive-form adjectives," in *Proceedings of Semantics and Linguistic Theory XXIII* (Ithaca, NY: CLC Publications), 587–610.

Lewis, D. (1970). General semantics. *Synthese* 22, 18–67. doi: 10.1007/BF00413598

Lidz, J., Halberda, J., Pietroski, P., and Hunter, T. (2011). Interface transparency and the psychosemantics of *most*. *Nat. Lang. Semant.* 6, 227–256. doi: 10.1007/s11050-010-9062-6

Link, G. (1983). "The logical analysis of plurals and mass terms: a lattice-theoretical approach," in *Meaning, Use and Interpretation of Language*, eds R. Bäuerle, C. Schwarze, and A. V. Stechow (Berlin: DeGruyter), 302–323.

Martin, R. M. (1958). *Truth & Denotation: A Study in Semantical Theory*. Routledge Library Editions: Epistemology. London; New York, NY: Routledge.

McNally, L. (2017). "On the scalar properties and telicity of degree achievements," in *Boundaries, Phases, and Interfaces*, eds O. Fernández-Soriano, E. Catroviejo, and Pérez-Jiménez (Amsterdam: John Benjamins), 174–192.

Moltmann, F. (2017). Natural language ontology. Oxford Research Encyclopedia of Linguistics.

Nakanishi, K. (2007). Measurement in the nominal and verbal domains. *Linguist. Philos.* 30, 235–276. doi: 10.1007/s10988-007-9016-8

Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Dev. Sci.* 21, 1–15. doi: 10.1111/desc.12533

Odic, D., Le Corre, M., and Halberda, J. (2015). Children's mappings between number words and the approximate number system. *Cognition* 138, 102–121. doi: 10.1016/j.cognition.2015.01.008

Odic, D., Pietroski, P., Hunter, T., Halberda, J., and Lidz, J. (2018). Individuals and non-individuals in cognition and semantics: the mass/count distinction and quantity representation. *Glossa* 3, 1–20. doi: 10.5334/gjgl.409

Parsons, T. (1979). "An analysis of mass terms and amount terms," in *Mass Terms: Some Philosophical Problems*, ed F. J. Pelletier (Dordrecht: D. Reidel Publishing Company), 137–166.

Partee, B. (2018). "Changing notions of linguistic competence in the history of formal semantics," in *The Science of Meaning: Essays on the Metatheory of Natural Language Semantics*, eds D. Ball and B. Rabern (Oxford, UK: Oxford University Press), 172–196.

Partee, B. H. (1980). "Montague grammar, mental representations, and reality," in *Philosophy and Grammar*, eds S. Kanger and S. Öhman (Dordrecht; Boston, MA: D. Reidel Publishing Company), 59–78.

Partee, B. H. (1995). "Chapter 11: Lexical semantics and compositionality," in *Invitation to Cognitive Science, Volume Part 1: Language, 2nd Edn.*, eds L. Gleitman and M. Liberman (Cambridge, MA: MIT Press), 311–360.

Pasternak, R. (2017). A lot of hatred and a ton of desire: Intensity in the mereology of mental states. *Linguist. Philos.* 42, 267–316. doi: 10.1007/s10988-018-9247-x

Pelletier, F. J. (2011). "Descriptive metaphysics, natural language metaphysics, Sapir-Whorf, and all that stuff: evidence from the mass-count distinction," *Baltic International Yearbook of Cognition, Logic, and Communication* (Manhattan, KS: New Prairie Press), 6.

Pietroski, P. (2010). Concepts, meanings, and truth: first nature, second nature, and hard work. *Mind Lang.* 25, 247–278. doi: 10.1111/j.1468-0017.2010.01389.x

Pietroski, P. (2018). *Conjoining Meanings: Semantics Without Truth Values*. Oxford, UK: Oxford University Press.

Pietroski, P., Lidz, J., Hunter, T., and Halberda, J. (2009). The meaning of *most*: semantics, numerosity, and psychology. *Mind Lang.* 24, 554–585. doi: 10.1111/j.1468-0017.2009.01374.x

Piñón, C. (2008). "Chapter 8: Aspectual composition with degrees," in *Adjectives and Adverbs: Syntax, Semantics and Discourse*, eds L. McNally and C. Kennedy (Oxford: Oxford University Press), 183–219.

Prasada, S., Ferenz, K., and Haskell, T. (2002). Conceiving of entities as objects and as stuff. *Cognition* 83, 141–165. doi: 10.1016/S0010-0277(01)00173-1

Putnam, H. (1975). "The Meaning of 'Meaning,"' in *Minnesota Studies in the Philosophy of Science, Volume III of* Scientific Explanation, Space, and Time, eds H. Feigl and G. Maxwell (Minneapolis, MN: University of Minnesota Press), 131–193.

Rips, L. J., and Hespos, S. J. (2015). Divisions of the physical world: concepts of objects and substances. *Psychol. Bull.* 141, 786–811. doi: 10.1037/bul0000011

Rotstein, C., and Winter, Y. (2004). Total adjectives vs. partial adjectives: scale structure and higher-order modifiers. *Nat. Lang. Semant.* 12, 259–288. doi: 10.1023/B:NALS.0000034517.56898.9a

Rullmann, H. (1995). *Maximality in the semantics of* wh*-constructions* (Ph.D. thesis). University of Massachusetts, Amherst, MA, United States.

Sassoon, G. (2011). "A *slightly* modified economy principle: stable properties have non-stable standards," in *Proceedings of the Israel Association of Theoretical Linguistics 27*, ed E. Cohen (Cambridge, MA: MITWPL), 163–181.

Schwarzschild, R. (2002). "The grammar of measurement," in *Proceedings of SALT XII*, ed B. Jackson (Ithaca, NY: CLC Publications; Cornell University), 225–245.

Schwarzschild, R. (2006). The role of dimensions in the syntax of noun phrases. *Syntax* 9, 67–110. doi: 10.1111/j.1467-9612.2006.00083.x

Schwarzschild, R. (2008). The semantics of comparatives and other degree constructions. *Lang. Linguist. Compass* 2, 308–331. doi: 10.1111/j.1749-818X.2007.00049.x

Schwarzschild, R. (2013). "Degrees and segments," in *Proceedings of Semantics and Linguistic Theory 23*, ed T. Snider (Ithaca, NY: Cornell University; CLC publications), 212–238.

Schwarzschild, R., and Wilkinson, K. (2002). Quantifiers in comparatives: a semantics of degree based on intervals. *Nat. Lang. Semant.* 10, 1–41. doi: 10.1023/A:1015545424775

Solt, S. (2012). "Comparison to arbitrary standards," in *Proceedings of Sinn und Bedeutung 16, Vol 16.2*, eds A. Aguilar Guevara, A. Chernilovskaya, and R. Nouwen (Cambridge, MA: MITWPL), 557–570.

Spelke, E. (2003). "What makes us smart? Core knowledge and natural language," in *Language in Mind: Advances in the Study of Language and Thought*, eds D. Gentner and S. Goldin-Meadow (Cambridge, MA: MIT Press), 277–311.

Spelke, E. S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behav. Dev.* 21, 181–200.

Strickland, B. (2016). Language reflects "core" cognition: A new theory about the origin of cross-linguistic regularities. *Cogn. Sci.* 41, 70–101. doi: 10.1111/cogs.12332

Taylor, B. (1977). Tense and continuity. *Linguist. Philos.* 1, 199–220.

Toledo, A., and Sassoon, G. W. (2011). "Absolute vs. relative adjectives - Variance within vs. between individuals," in *Semantics and Linguistic Theory, Vol. XXI*, eds N. Ashton, A. Chereches, D. and Lutz (Ithaca, NY: CLC Publications), 135–154.

Tomaszewicz, B. (2011). "Verification strategies for two majority quantifiers in Polish," in *Proceedings of Sinn und Bedeutung 15*, ed I. E. A. Reich (Saarbrücken: Saarland University Press).

Travis, C. (1996). Meaning's role in truth. *Mind* 105, 451–466.

Unger, P. (1971). A defense of skepticism. *Philos. Rev.* 80. 198–219.

Vicente, A. (2012). On Travis cases. *Linguist. Philos.* 35, 3–19. doi: 10.1093/mind/105.419.451

von Stechow, A. (1984). Comparing semantic theories of comparison. *J. Semant.* 3, 1–77. doi: 10.1093/jos/3.1-2.1

Wellwood, A. (2012). "Back to basics: *more* is always *much-er*," in *Proceedings of Sinn und Bedeutung 17*, eds E. Chemla, V. Homer, and G. Winterstein (Cambridge, MA: MITWPL).

Wellwood, A. (2015). On the semantics of comparison across categories. *Linguist. Philos.* 38, 67–101. doi: 10.1007/s10988-015-9165-0

Wellwood, A. (2018). "Structure preservation in comparatives," in *Semantics and Linguistic Theory (SALT) 28*, eds S. Maspong, B. Stefánsdóttir, K. Blake, and F. Davis (Ithaca, NY: CLC Publications), 78–99.

Wellwood, A. (2019). *The Meaning of* More. Studies in Semantics and Pragmatics. Oxford: Oxford University Press.

Wellwood, A., Gagliardi, A., and Lidz, J. (2016). Syntactic and lexical inference in the acquisition of novel superlatives. *Lang. Learn. Dev.* 12, 262–279. doi: 10.1080/15475441.2015.1052878

Wellwood, A., Hacquard, V., and Pancheva, R. (2012). Measuring and comparing individuals and events. *J. Semant.* 29, 207–228. doi: 10.1093/jos/ffr006

Wellwood, A., He, A. X., Lidz, J., and Williams, A. (2015). "Participant structure in event perception: towards the acquisition of implicitly 3-place predicates," *Univ. Pennsyl. Work. Pap. Linguist.* 21, 1–9.

Wellwood, A., Hespos, S. J., and Rips, L. (2018a). How similar are objects and events? *Acta Linguist. Acad.* 15, 473–501. doi: 10.1556/2062.2018.65.2-3.9

Wellwood, A., Hespos, S. J., and Rips, L. (2018b). "Chapter 8: The *object : substance :: event : process* analogy," in *Oxford Studies in Experimental Philosophy, Vol. II*, eds T. Lombrozo, J. Knobe, and S. Nicholas (Oxford: Oxford University Press), 183–212.

Williams, A. (2015). *Arguments in Syntax and Semantics, Volume, Key Topics in Syntax*. Cambridge, UK: Cambridge University Press.

Conflict of Interest: 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.

frontiers in Psychology

METHODS
published: 31 January 2020
doi: 10.3389/fpsyg.2020.00073

The Application of Signal Detection Theory to Acceptability Judgments

Yujing Huang and Fernanda Ferreira*

Department of Psychology, University of California, Davis, Davis, CA, United States

Acceptability judgments have been an important tool in language research. By asking a native speaker whether a linguistic token is acceptable, linguists and psycholinguists can collect negative evidence and directly test predictions by linguistic and psycholinguistic theories, which provide important insight into the human language capacity. In this paper, we first give a brief overview of this method including: (1) the linking hypothesis for this method, (2) the controversy about the test, and (3) limitations of the current analysis of the results. Then, we propose a new way of analyzing the data: Signal Detection Theory. Signal Detection Theory has been used in many other psychological research areas such as recognition memory and clinical assessments. In this paper, using two examples, we show how Signal Detection Theory can be applied to judgment data. The benefits of this approach are that it can: (1) show how well participants can differentiate the acceptable sentences from unacceptable ones and (2) describe the participant's bias in the judgment. We conclude with a discussion of remaining questions and future directions.

Keywords: signal detection theory, acceptability judgments, d-prime, response bias, one-factor design, two-factor design

OPEN ACCESS

Edited by:
Viviane Marie Deprez, Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Evelina Leivada, University of Rovira i Virgili, Spain
Asya Achimova, University of Tübingen, Germany

****Correspondence:***
Yujing Huang
yujhuang@ucdavis.edu

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

Received: *15 September 2019*
Accepted: *10 January 2020*
Published: *31 January 2020*

Citation:
Huang Y and Ferreira F (2020) The Application of Signal Detection Theory to Acceptability Judgments. Front. Psychol. 11:73. doi: 10.3389/fpsyg.2020.00073

A BRIEF OVERVIEW OF ACCEPTABILITY JUDGMENTS

One important type of linguistic data comes from judgments of the well-formedness of linguistic stimuli. An early justification for the use of judgments comes from Chomsky (1957, p. 13), in which it is stated that "[t]he fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences." In this view of language research, grammar is not a set of rules which passively describe what has been seen in a language, but can be viewed as a system for evaluating sequences and making clear predictions regarding what is allowed or disallowed in a language. This makes the linguistic theory falsifiable. Different from methods such as corpus analysis, which can show what structures are possible in a language, linguistic judgments may also reveal what structures are disallowed. These judgments therefore provide negative evidence and allow researchers to directly test predictions regarding what forms a grammar generates and which it does not. Compared to observational data which should not be altered, linguistic judgments can be elicited to target specific hypothesis in a systematic manner.

When judgments were first collected to elicit linguistic intuitions, the procedure was quite informal. These took the form of grammatical judgments. To collect grammatical judgments, linguists would ask their fellow linguist to judge whether a sentence is grammatical or not. Based on this judgment, they would conclude whether a grammatical principle was supported or falsified. The reason the procedure involved querying fellow linguists is because a linguist is tuned to detect subtle grammatical differences and can separate syntactic factors from other

influences such as semantics and pragmatics. However, this informal procedure has several potential problems. First, the judgment is based on very limited stimuli which can be as few as one or two token examples (e.g., *Who do you think that left* for the so-called "that-trace" effect; Perlmutter, 1968). Validating a grammatical principle with such a limited sample can be problematic because the generalizability of the judgment across different items is unknown. Second, there may be some implicit bias in the judgment because linguists' judgments may be unconsciously influenced by the theory they know. Third, it is unclear whether the judgment from a single person can be generalized to the entire population. Fourth, without a standard procedure, the stimuli could be created with different standards by different linguists. Some linguists may compare only minimal pairs. For example, when comparing the well-formedness of prenominal modifiers of different verbs, they may test *the fallen boy* compared to *the jumped boy*, changing only the critical past participle in the sentence. Others may compare *the fallen leaf* with *the laughed boy*. Changing the noun in the phrase could introduce potential confounds. Because of these problems, the reliability of grammatical judgments elicited as described here has been questioned (Schütze, 1996; Edelman and Christiansen, 2003; Wasow and Arnold, 2005; Culicover and Jackendoff, 2010; Gibson and Fedorenko, 2010).

To increase reliability, some researchers advocate using formal procedures that are standardly used in psychology to collect linguistic judgment data (Schütze, 1996; Ferreira, 2005; Culicover and Jackendoff, 2010; Gibson and Fedorenko, 2010; i.a.). In the formal procedure, there need to be multiple items for the same condition with careful controls for potential confounds, and the data are usually collected from several naïve participants who have limited to no exposure to linguistic theory. This formal procedure will increase the sample size of participants and items, will better control for confounds, and avoids bias based on adherence to a particular linguistic theory. While the reliability of the informal procedure has been much debated (Gibson and Fedorenko, 2010; Sprouse and Almeida, 2012; Gibson et al., 2013a,b), it has been shown that acceptability judgments are generally reliable when formal data collection procedures are used that conform to the standards of experimental psychology (Langsford et al., 2018; Linzen and Oseki, 2018). Therefore, in this paper, we restrict our discussion to formal data collection procedures.

However, it is a misnomer to call the data collected using these experimental standards for collecting data as grammaticality judgments. From a theoretical perspective, naïve participants may not be able to separate syntactic factors from other factors such as frequency and plausibility. Their judgments are not based solely on the grammaticality of the stimuli. From a practical perspective, if the participants are asked to judge grammaticality, they are likely to judge the stimuli based on the prescriptive grammar they learned in school rather than providing their intuition about the well-formedness of the stimuli. The better practice may be to ask participants about the acceptability of the stimuli rather than their grammaticality. In asking participants about acceptability, the judgments may be influenced by factors other than the grammaticality of the stimuli, such as frequency, plausibility, pragmatics as well as processing difficulty and processing accuracy. Therefore, it is more appropriate to refer to these judgments as acceptability judgments.

Acceptability judgments differ from grammaticality judgments in an important way: grammaticality reflects the nature of the linguistic stimuli while "acceptability is a percept that arises (spontaneously) in response to linguistic stimuli that closely resemble sentences" (Schütze and Sprouse, 2014). On this view, acceptability is no different from other percepts such as loudness or luminance. One important feature of human perception is that it is never perfect. There is always noise in the perceptual data and in perceptual systems. Indeed, if we ask the same participant to judge different items in the same condition or if we ask different participants to judge the same item, we would not necessarily expect the same response from every participant on every item. If we look at the results from studies that test the reliability of acceptability judgments, we can see that there is indeed between-subject and between-item variability (e.g., Langsford et al., 2018).

This noise can come from many different sources. As we mentioned above, many factors can influence the perception of the acceptability of a sentence, for example, plausibility, frequency, etc. If the event described in a sentence is less plausible, a participant may judge it to be less acceptable although the sentence is perfectly grammatical. Such factors are based on participants' unique linguistic and nonlinguistic experiences and differ from person to person. They can be controlled as a whole with norming studies but are hard to eliminate for individual participants. As a result, there will be variability in judgments at individual participant and individual item level. In addition, processing difficulty can also influence the acceptability of a linguistic stimulus. For example, a garden-path sentence such as "The horse raced past the barn fell" may be judged as unacceptable although it is not ill-formed. This is because this sentence is hard to parse and the participant may have a hard time building the correct representations of the sentence and therefore will interpret difficulty of processing as evidence for ungrammaticality (Ferreira and Henderson, 1991). Finally, as Gibson et al. (2013a) have argued, input to our language processing mechanisms is not error-free. A participant could provide a judgment based on an input that is not entirely consistent with the stimuli. For example, a participant may misread a sentence because the form of a sentence does not conform to his/her predicted structure and judge an ungrammatical sentence to be grammatical as a result. These are inherent features of our language processing mechanisms and cannot be eliminated either. As none of these sources of noise can be eliminated, there will always be some variance in acceptability judgments.

Another important feature of perception is that there can be some biases in the response. In cases when the stimuli are entirely unacceptable, bias may not be a concern; presumably, nobody will judge a random sequence of words as acceptable, for example. However, in less clear cases, the response bias may have impact on the data. Some participants may be reluctant to judge a sentence as unacceptable and therefore will have a bias to say *yes*. Other participants may tend to be very strict and judge anything that sounds a bit odd to them to be unacceptable (no matter whether it is the form of the

sentence, the plausibility of the scenario, or other reasons). These participants have a bias to say *no*. These biases can reduce the difference between theoretically unacceptable and acceptable stimuli and therefore need to be taken into consideration in the data analysis models.

Acceptability judgment data are usually analyzed using some type of significance test, for example, *t*-test (e.g., Clifton et al., 2006; Sprouse, 2011; Sprouse et al., 2013; i.a.) and mixed effect models (e.g., Gibson et al., 2013b; Sprouse et al., 2013, i.a.)[1]. With these tests, a single value of *p* would tell us whether we should reject the null hypothesis and adopt the alternative hypothesis, i.e., these two samples are significantly different from each other. Because these tests compare two samples, some variability is assumed in the data. Therefore, noise is not a problem for these models.

However, these significance tests do not have a built-in mechanism to model response biases. *T*-tests which care about the sample means could be impacted by the bias because the bias may dilute the differences between the two samples. Mixed-effect models can capture the variability at the participant level if a participant random effect is added to the statistical model, but this is still different from modeling response bias[2]. Response biases are not merely random variability across participants. Instead, they are systematic and reflect the criterion a participant sets, i.e., the threshold to judge a stimulus as acceptable. The information of the criterion is overlooked in these significance tests.

In addition to the inability to model biases, there is another factor we need to consider regarding the use of significance tests to evaluate judgment data: How should we interpret any significant results from these models? For example, if the mean of one condition is 0.5 and another is 0.6, given a large sample size, it is likely that a significance test would give a value of *p* that is below our predetermined alpha-level (say, 0.05). Does this significant result mean anything? We could easily run into the standard caveat of significance testing, i.e., the statistical significance may not be meaningful given our theory. One solution to solve this problem is to calculate the effect size. This can be straightforward with the *t*-test but quite complex in mixed-effect models which are more appropriate for tests with multiple subjects and items (Westfall et al., 2014; Brysbaert and Stevens, 2018).

In this section, we gave a brief overview of acceptability judgment in language research. We discussed the linking hypothesis for using acceptability judgments to study language and we also briefly reviewed the nature of judgment data. In the remainder of this paper, we discuss an alternative method of analyzing the acceptability judgment data, i.e., signal detection theory, which models the size of the effect directly and offers a straightforward measure of bias. In the section "Signal Detection Theory and Acceptability Judgments," we explain SDT and how it can help us better understand the acceptability judgment data. In the sections "Signal Detection Theory and One-Factor-Design Experiments" and "Signal Detection Theory and Two-Factor-Design Experiments," we provide two examples of the application of SDT to acceptability judgment. And in the final section, "Discussion and Future Directions," we discuss some remaining questions and future directions.

SIGNAL DETECTION THEORY AND ACCEPTABILITY JUDGMENTS

Signal Detection Theory (SDT) was originally designed to describe the ability of an observer to decide whether the source of a voltage change is noise or signal plus noise (Peterson et al., 1954). Soon afterward, it was adopted by cognitive scientists to measure human decision making in perceptual studies (Tanner and Swets, 1954; Swets et al., 1961). SDT assumes that performance is not perfect and describes how well observers can discriminate or recognize certain signals given the background noise. For example, in recognition memory studies, participants need to decide if a specific stimulus has been presented or not (old or new). There is some ambiguity in this decision, so that given the same stimulus, a participant may judge it as either old or new. SDT captures sensitivity in discriminatory ability so that higher sensitivity means the participant is better able to discriminate old from new items.

SDT has also been adopted in language research by psychologists and linguists to investigate speech perception. In speech perception studies, participants may be asked to categorize sounds according to whether they belong to a certain category or if two sounds are different from each other, corresponding to two commonly used paradigms, "yes-no" and "ABX." In a study making use of the "yes-no" paradigm, participants decide whether a single signal "A" is present. In the "ABX" paradigm, the two sounds being discriminated ("A" and "B") are followed by a repetition of one of them, and participants are asked to decide whether "X" is the same as "A" or "B." Participants' ability to discriminate the sounds is described by a sensitivity measure. In the design, the stimuli "A" and "B" can be a fixed standard or "roving" on a continuum. Participants' strategy may change accordingly: With a fixed standard, they may first categorize the stimulus and then compare the categories, and with a "roving" standard, participants may apply a threshold to compare the stimuli and decide if they are different enough to be labeled as such. With different strategies, the calculation of discrimination sensitivity also may differ (Macmillan et al., 1977; Macmillan and Creelman, 2004).

It has been argued that acceptability judgments are a reported perception of acceptability (Chomsky, 1965; Schütze, 1996;

[1]An anonymous reviewer pointed out that in the discussion of statistical methods, one method that is worth mentioning is Bayesian statistics. Bayesian statistics provides a probability distribution over hypotheses. It can be especially useful when we want to integrate prior beliefs into the analysis. However, it shares some limitation with frequentist tests when modeling perpetual data (e.g., it does not have an explicit way to quantify bias).

[2]An anonymous reviewer pointed out that a random intercept can provide some insights on bias by showing that "the acceptability judgment value never goes below a certain threshold for a given subject." However, there are two limitations with this random intercept argument. First, the inference concerning bias is indirect (we need to compare the intercept with some value that must be separately calculated). Second, when the subject effect is treated as a random effect, it is essentially seen as variance that researchers do not care about (compared to a main effect). However, bias is not random noise: as we discuss in this paragraph, bias reflects the decision criterion of a participant.

Sprouse and Almeida, 2012). In acceptability judgment studies, participants receive a sensory input in the form of a linguistic sequence and are asked to decide whether the sequence is acceptable. This is similar to perceptual studies in other domains, for example, recognition memory studies mentioned above. The SDT was previously adopted by Achimova (2014) to analyze acceptability data related to quantifier scope but the work does not discuss why SDT is appropriate for judgment data, nor does it mention how the different metrics were calculated. In this section, we show why SDT is appropriate for analyzing acceptability judgments and we describe some advantages of using this method as well as different measures in SDT.

As was discussed in the section "Signal Detection Theory and Acceptability Judgments," acceptability judgments assume a single underlying construct, i.e., acceptability. Participants need to make a decision regarding this construct: whether a sentence is acceptable or not[3]. For a single category, there is a probability distribution of judgments along the dimension of this construct. As there are two categories of stimuli, acceptable and unacceptable, there are two probability distributions that differ from each other. If we use the x-axis to represent the rating of the items and the height to represent the probability of the rating, we will see two distributions similar to those in **Figure 1**. Because there is some noise in decision making (participants may not always be able to tell if a sentence is acceptable or not due to various sources of noise), there is an overlapping area in these two distributions. In **Figure 1**, for example, an item that receives an average rating of 0.2 is likely to be an unacceptable item whereas an item that receives an average rating of 0.8 is likely to be an acceptable item. If an item receives an average rating of 0.5, it is equally likely to be an acceptable or unacceptable item.

Instead of focusing on the distributions of the ratings as significant tests usually do, SDT evaluates the type of decision being made. From the perspective of signal detection theory, in an acceptability judgment experiment, there are two types of stimuli and two possible decisions[4]. This creates four logical combinations. If the stimulus is predicted as acceptable by a linguistic theory and is judged as acceptable, it is a *hit* (i.e., true positive). If the stimulus is predicted as acceptable by a linguistic theory and judged unacceptable, it is a *miss* (i.e., false negatives). If the stimulus is predicted as unacceptable by a linguistic theory but judged as acceptable, it is a *false alarm* (i.e., false positives). If the stimulus is predicted as unacceptable by a linguistic theory and judged as unacceptable, it is a *correct rejection* (i.e., true negative). There are thus two types of correct responses and two types of errors. **Table 1** is a summary of these four types of outcomes.

After categorizing the responses, we can calculate the likelihood ratio of each category. For example, the hit rate (H) is the proportion of acceptable trials to which the participant responded "acceptable." False alarm rate (F) is the proportion of unacceptable trials to which the participant responded "acceptable." Assuming that *hit* is 20, *false alarm* is 10, *miss* is 5, and *correct rejection* is 15 (see **Table 2**), hit rate is 20/(20 + 5) = 0.8 and false alarm rate is (10/10 + 15) = 0.4.

$$d' = z(H) - z(F)$$

The measure of participants' ability to distinguish between the stimuli (sensitivity, d') is defined by the inverse of the normal distribution function of H and F (Green and Swets, 1966). In the example above, z(H) is 0.842, z(F) is −0.253, and d' is z(H) − z(F) which is equal to 1.095. The sensitivity reflects the distance between the acceptable and unacceptable distributions (**Figure 2**). The larger this number is, the higher the sensitivity (the more distant the two distributions).

[3]An anonymous reviewer has pointed out that acceptability is gradient rather than binary. To clarify, when we talk about binary decisions, we refer to the nature of the task (i.e., in the judgment study, the participants are asked to judge if a stimulus is acceptable). This does not require the underlying construct to be binary. To make binary judgments on a continuous underlying construct, the participant must decide on a threshold beyond which all the stimuli are acceptable and below which all the stimuli are unacceptable. This is how a continuous underlying construct can be measured with a binary decision. This follows the same logic as tasks in memory research in which the participant judges the familiarity of the stimuli (continuous) by providing binary judgments (if the stimuli have been seen before).

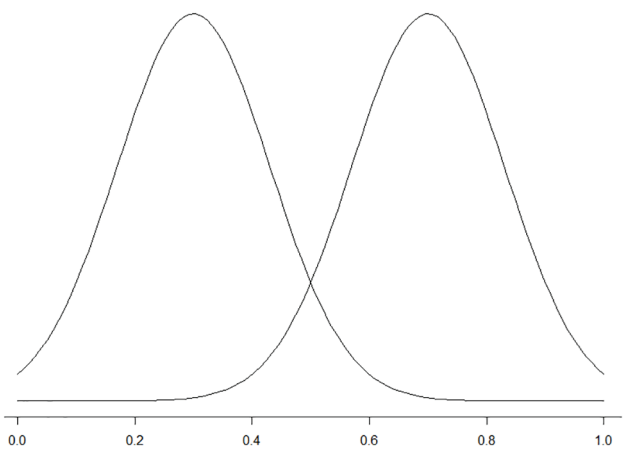


FIGURE 1 | Visual illustration of the probability distributions.

[4]Bader and Häussler (2010) show that gradient judgment data and binary judgment data are highly correlated. Therefore, in this paper, we adopt the binary task which makes the data structure simple and straightforward. An anonymous reviewer pointed out that "there are self-evident judgments, whose replication/correlation across different elicitation techniques is unsurprising, and then there are potentially questionable judgments, which may introduce some variation across techniques/samples". In Bader and Häussler (2010), many different phenomena were tested and in their results, it is clear that the judgments are not polarized (which is what we would expect if the sentences are cleared acceptable or unacceptable). Therefore, Bader and Häussler (2010) did not only test self-evident judgments.

TABLE 1 | Categories of judgments based on SDT.

		Signal	
		Acceptable	**Unacceptable**
Response	**Acceptable**	Hit	False alarm
	Unacceptable	Miss	Correct rejection

In addition to measuring participants' sensitivity with respect to discriminating the two sets of stimuli, we can also quantify the bias of participants. Bias is caused by participants' tendency to give one type of response, either "yes" or "no." As we discussed in the section "Signal Detection Theory and Acceptability Judgments," if a participant is reluctant to say any sentence is unacceptable, that participant has a "yes" bias; if a participant tends to say any sentence is unacceptable, that participant has a "no" bias. There are many different ways to quantify bias, for example, criterion location (c), relative criterion location (c'), and likelihood ratio (beta). The comparison among these three indices is too technical and beyond the scope of this paper. Here, we use criterion location (c) for illustration purpose. This is because this measure depends monotonically on H and F in the same direction and it is independent of sensitivity d' (Stanislaw and Todorov, 1999; McNicol, 2005). However, whether it is the best measure of bias for acceptability judgment is an empirical question that needs further investigation.

$$c = -\frac{1}{2}\left(z(H) + z(F)\right)$$

Criterion location is defined as the negative value of half of the sum of z(H) and z(F). Conceptually, it describes the distance between the selection criterion (the threshold for giving a certain type of response) and the midpoint of the two distributions. When the false alarm and miss rates are equal, c equals 0; when false alarm rate is smaller than misses, c is positive and vice versa. For example, in **Figure 3**, the threshold is set to 0.2. Any rating higher than 0.2 is judged acceptable and anything lower than 0.2 is judged unacceptable. If the left curve represents unacceptable stimuli and the right curve represents acceptable stimuli, the area A1 (the red shaded area) represents the probability of the correct rejection, A2 (the blue shaded area) represents the probability of the false alarms, A3 (the green shaded area) represents the probability of miss, and A4 (the gray shaded area) represents the probability of hits. In **Figure 3**, the false alarm area is larger than the misses (A1 > A3), and the bias is negative. This means that the participant has a "yes" bias (is more likely to judge the stimuli as acceptable rather than unacceptable regardless of the properties of the stimuli). In the example of **Table 2**, c is −0.294. That is a "yes" bias.

SDT is not merely an alternative statistical analysis to acceptability judgment data. It is a different way to think about acceptability judgments. Significance tests assess whether the two samples tested are from the same underlying distribution. This may create an illusion that we are testing the nature of the linguistic stimuli, that is, whether the stimuli are acceptable or not. However, acceptability is not a reflection on the nature of the stimuli. Rather, it reflects how these stimuli are perceived. Therefore, what is tested should not be whether these two sets of stimuli come from the same underlying distribution. Rather, the question should be whether the two sets of stimuli are perceptually differently. SDT is designed to address the latter while significant tests address the former.

SIGNAL DETECTION THEORY AND ONE-FACTOR-DESIGN EXPERIMENTS

In this section, we provide a concrete example of the application of SDT to acceptability judgments with a one-factor design. The data are taken from a study in Huang (2018)[5]. The aim of the study was to investigate one of the unaccusative diagnostics – the -er nominalization (nominalizing a verb by adding the -er morpheme, e.g., run - > runner). The Unaccusative Hypothesis

TABLE 2 | A toy example of judgment data with number of participant responses in each of the four categories defined by the signal detection analysis.

Hit	False alarm
(20)	(10)
Miss	Correct rejections
(5)	(15)

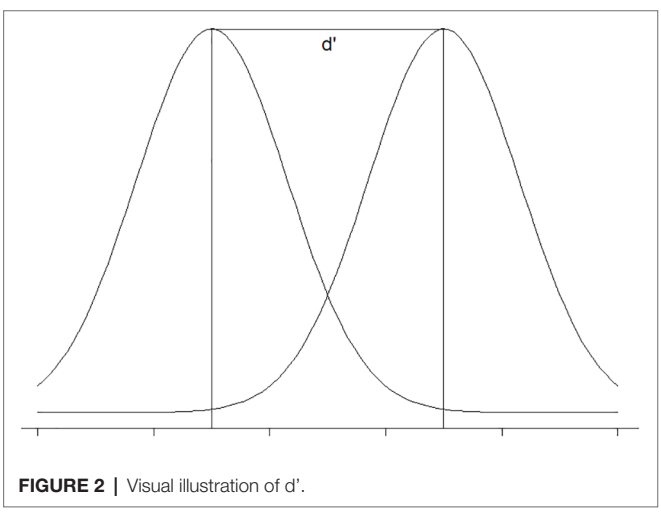


FIGURE 2 | Visual illustration of d'.

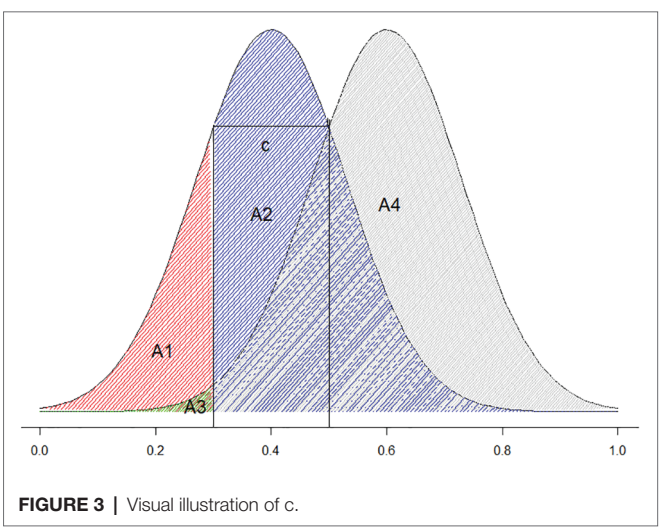


FIGURE 3 | Visual illustration of c.

[5]The description in this section is an oversimplification of the actual study. The original data were based on a 7-point scale. We split the data into binary choices at the midpoint (any score below 4 is transformed to 0 and the rest is 1). We only took a subset of the data for illustration purpose. We do not intend to make any theoretical conclusion.

claims that there are two types of intransitive verbs. The subject of the unergative verb (e.g., *run*) is base-generated as the external argument, whereas the subject of the unergative verb (e.g., *arrive*) is originally generated as the internal argument (Perlmutter, 1978; Maling et al., 1986). Fabb (1984) has argued that -er nominalization only applies to a verb that has an external argument. Therefore, −er nominalizations should be possible for unergative verbs and not unaccusative verbs. Based on the theory, we can construct a study to test if English speakers can distinguish unaccusative verbs and unergative verbs using -er nominalizations. In Huang (2018), each participant was given a list of unaccusative and unergative verbs with the -er nominalization (e.g., *runner* versus *arriver*, where presumably *arriver* seems unacceptable) and was asked to judge if the word was an acceptable English word. For the purposes of this exercise, we use a subset of the data only. In this subset, there were 30 unaccusative verbs and 30 unergative verbs with an -er nominalization. All the items were judged by 20 native English speakers who were naive with respect to the linguistic and psycholinguistic theories. Unaccusativity of the verb was the only factor manipulated in the study, and it had two levels: unaccusative and not unaccusative (i.e., unergative).

Overall Sensitivity and Bias

To assess whether the unaccusative and unergative conditions are perceived differently, we can calculate the overall sensitivity and bias based on the collective judgments. This means that we ignore individual differences across items and participants. To calculate sensitivity and bias, first we need the frequency of each type of judgment. Those frequencies are given in **Table 3**.

As we explained above, the unergative condition should be judged as acceptable and therefore, the acceptable responses are *hits* and the unacceptable responses are *misses*. There are 526 *hits* and 74 *misses*. The unaccusative condition should be judged unacceptable and therefore the acceptable responses are *false alarms* and the unacceptable responses are *correct rejections*. There are thus 331 *correct rejections* and 269 *false alarms*. The data are summarized in **Table 4**.

As shown in the section "Signal Detection Theory and One-Factor-Design Experiments," hit rate (H) is Hit/(Hit+Miss) which is 526/(526 + 74) = 0.88. False alarm rate (F) is False alarm/(False alarm+Correct rejection) which is 269/(269 + 331) = 0.45. Based on the hit rate and the false alarm rate, we can calculate d' (sensitivity) and c (bias). The sensitivity d' is z(H) − z(F) = 1.158 − (−0.130) = 1.288. The bias c is −½(z(H) + z(F)) = −0.5*(1.158–0.130) = −0.514. In the context of the study, the value of d' is the distance between the unaccusative and unergative distributions, which is 1.288. This is a non-zero value, meaning that participants were able to discriminate unaccusative and unergative stimuli (the perceptual distance between the unaccusative and unergative stimuli is not zero). The negative bias means that the participants (as a whole) have a bias to judge the stimuli as acceptable.

However, before we reach any strong conclusion, we would want to ask if the d' and bias we estimated from our data reflect the true underlying parameters. Gourevitch and Galanter (1967) provided a way to calculate the variance of d' and c by using an approximation. The variance of d' can be calculated by the equation below:

$$var(d') = \frac{H(1-H)}{N_2\left[\Phi(H)\right]^2} + \frac{F(1-F)}{N_1\left[\Phi(F)\right]^2}$$

where N_2 is the number of signal trials and N_1 is the number of noise trials. Φ(H) is the height of the normal density function at z(H).

As we have calculated, H is 1.158 and F is −0.130. Based on the equation above, Φ(H) is 0.204 and Φ(F) is 0.396. Var(d') is 0.00697. The standard error is the square root of the variance: 0.083. The 95% confidence interval is 1.96 standard errors above and below the estimated d' and therefore is 1.288 ± 1.96*0.083, that is (1.12, 1.45). This means that we can be 95% confident that the true d' is between 1.12 and 1.45. Critically, this interval does not contain 0. Therefore, the participants were able to discriminate the unaccusative stimuli from the unergative stimuli in the study based on the nominalization test.

The variance of bias is a quarter of the variance of d' (Macmillan and Creelman, 2004). Therefore, the variance of c is 0.0017, the standard error is 0.042 and the confidence interval is −0.514 ± 1.96*0.042, which is (−0.68, −0.35). This interval is negative and, therefore, there is a bias to judge the stimuli as acceptable.

Sensitivity and Bias by Participant

In recognition memory research (for an overview of such work, see Rugg and Curran, 2007), sensitivity and bias are usually calculated at each individual participant level. This is because sensitivity and bias describe the perceptions of individual participants and can differ from person to person. Some people may be better at discriminating certain stimuli than others and some people may tend to say "yes" or "no" more than others.

As we discussed in the section "Signal Detection Theory and Acceptability Judgments," individual linguistic and non-linguistic experiences differ from person to person. Therefore, their judgment of the stimuli can differ from individual to individual. If we want to make a claim about an entire population (e.g., American English speakers), we need to test the hypothesis at the individual level and see if the hypothesis holds across individuals. This is the first step to making any generalization about the population.

TABLE 3 | Frequency of the choices in each category for the –er nominalization study.

	Unergative	Unaccusative
Acceptable	526	269
Unacceptable	74	331

TABLE 4 | Number of participant responses in each of the four categories defined by the signal detection analysis for the -er nominalization study.

Hit	False alarm
(526)	(269)
Miss	Correct rejections
(74)	(331)

The steps to calculate individual sensitivity (d') and bias (c) are the same as those for the overall d' and c. Instead of summarizing the data across all participants, we categorize and summarize the responses by each individual. In our example, there were 30 trials in each condition. It is possible that a participant will have perfect accuracy (hit rate equals 1). This would result in an infinite d'. There are two common ways to correct for extreme proportions. One is to add 0.5 to all data cells for that participant. The other is to convert proportion of 0 to 1/(2 N) and 1 to 1–1/(2 N), where N is the number of trials. Here, we choose to add 0.5 to all data cells. This method is proved to be less biased and more conservative (Hautus, 1995).

After calculating the sensitivity and bias for each participant, we can perform inferential statistics on each. Because our question is whether participants can discriminate the two conditions, we want to know if the perceptual distance (d') is likely to be 0. To answer this question, we can perform a one sample *t*-test to test if 0 is a likely d' value based on our sample. We found that our sample mean is significantly different from 0 ($t = 13.19$, $p < 0.001$). Therefore, our participants were able to discriminate unaccusative and unergative stimuli.

Following the same logic, we can run a *t*-test and see if the bias is different from 0 (no bias). We find that the bias significantly different from 0 ($t = -5.73$, $p < 0.001$).

SIGNAL DETECTION THEORY AND TWO-FACTOR-DESIGN EXPERIMENTS

In section Signal Detection Theory and One-Factor-Design Experiments, we gave an example of how SDT can work with one-factor-design experiments. In this section, we show how SDT can be applied to two-factor-design studies. The data in this section are taken from another study in Huang (2018). This study investigated another unaccusative diagnostic: prenominal participles. Prenominal participles refer to the phenomenon where the participle form of a verb can be used as a prenominal modifier of a noun (e.g., *fallen* in *the fallen leaf*). It has been argued that prenominal participles are only possible when the verb is unaccusative and impossible when the verb is unergative (Borer, 1984; Levin and Rappaport, 1986). In Huang (2018), these claims were tested using acceptability judgments[6]. In this study, there were two types of verbs (unaccusative and unergative) and two conditions (control and test). The test condition was a noun phrase with the prenominal modifier (e.g., *the fallen leaf*) and the control condition was a sentence in which the verb was the predicate and the noun was the argument (*The leaf fell.*). Each verb appeared in both the test and control conditions. The control condition was added to ensure that the combination of the verb and the noun was not semantically or pragmatically unacceptable. Two lists of stimuli were created so that each participant only saw the same verb once. The study used a counterbalanced design. The data analyzed in this paper came from 18 participants in each list resulting in a total of 36 participants. There were 30 unaccusative and 30 unergative verbs.

Overall Sensitivity and Bias

Similar to the previous section, we can calculate the overall sensitivity and bias across all the participants and items. These metrics will tell us whether the participants discriminated unaccusative and unergative stimuli as a whole and whether there is evidence of bias in their responses. Different from the study described in the section "Signal Detection Theory and One-Factor-Design Experiments," the current study followed a 2 × 2 design. In addition to the verb factor, we added a condition factor where a verb appeared in both the test and control conditions. We do not expect the judgment patterns to be the same in the test and control conditions. In fact, if the prenominal participle test can differentiate unaccusative verbs from unergative verbs, we would expect participants to discriminate the two types of verbs in the test condition but not in the control condition (because the control condition does not have prenominal modifiers and is therefore acceptable for both verb types). Thus, we need to analyze these two conditions separately.

For the test condition, the number of acceptable and unacceptable judgments for the two verb types is summarized in **Table 5**.

As we explained above, the unaccusative condition should be judged as acceptable and therefore the acceptable responses are *hits* and the unacceptable responses are *misses*. There are 285 *hits* and 255 *misses*. The unergative condition should be judged unacceptable and therefore the acceptable responses are *false alarms* and the unacceptable responses are *correct rejections*. There are 118 *false alarms* and 422 *correct rejections*. The data are summarized in **Table 6**.

Based on **Table 6**, d' for the test condition is 0.847 and c is 0.354. In the context of the study, the value of d' is the distance between the unaccusative and unergative distributions, which is 0.847. This is a non-zero value, meaning that the participants can discriminate unaccusative and unergative stimuli (the perceptual distance between the unaccusative and unergative stimuli is not zero). The positive bias means that the participants (as a whole) have a bias to judge the stimuli as unacceptable.

TABLE 5 | Frequency of the choices in the test condition for the prenominal participle study.

	Unaccusative	Unergative
Acceptable	285	118
Unacceptable	255	422

TABLE 6 | Number of participant responses in each of the four categories defined by the signal detection analysis for the prenominal participle study.

Hit	False alarm
(285)	(118)
Miss	Correct rejections
(255)	(422)

[6]This is again an oversimplification of the study. The counterbalanced structure was also altered to work with SDT. We do not intend to make any theoretical conclusion with this example. All interpretations of the data are for illustration purposes to show what d' and c mean in a real dataset.

As in the section "Signal Detection Theory and Two-Factor-Design Experiments," we can calculate the standard error and 95% confidence interval of d'. The standard error is 0.0809 and the confidence interval is (0.69, 1.01). This interval does not contain zero which means that there is a non-zero perceptual distance between unaccusative and unergative stimuli. In other words, the participants were able to discriminate these two sets of stimuli.

Following the same steps, we can also calculate d' and c in the control condition. **Table 7** summarizes the frequency of responses.

One thing to note is that the categorization of the control condition is artificial, because all control sentences should be judged as acceptable no matter what type of verb they include. However, when we analyze the data, we need to categorize the responses in the same way as in the test condition so that the interpretation of d' and c remains the same and can be compared across test and control conditions. If an unaccusative stimulus is judged as acceptable, it is a *hit* and otherwise it is a *miss*. There are 285 *hits* and 255 *misses*. Likewise, if an unergative stimulus is judged as unacceptable, it is a *correct rejection*, and otherwise it is a *false alarm*. There are 422 *correct rejections* and 118 *false alarms*. The data are summarized in **Table 8**. In hypothesis tests such as the *t*-test, we assume that the null hypothesis is true and test if we should reject this assumption. Here, we assume that the two distributions of interest can be discriminated (the unaccusative stimuli should be acceptable and unergative stimuli should be unacceptable) and test whether this is true.

Based on **Table 8**, the control condition has a d' of −0.156 and a c of −1.623. The standard error of d' is 0.127 and the 95% confidence interval is (−0.41, 0.09). This confidence interval contains 0. Therefore, we have no evidence that the participants discriminated the unaccusative and unergative stimuli in the control condition. This is consistent with our expectations, since the verb+noun sequence was predicted to be acceptable for both verb types. There is no theoretical reason why these two sets of stimuli would differ in the control condition.

Taken together, the results show that participants were able to discriminate unaccusative and unergative verbs in the prenominal participle form, and this ability is not confounded with any semantic and pragmatic differences, since the verbs were not distinguished in the control condition. The calculation of confidence interval for c is the same as that in the one-factor design section and so we will not repeat it here.

Sensitivity and Bias by Participant

The calculations of sensitivity and bias by participant are very similar to those of the section Signal Detection Theory and One-Factor-Design Experiments. The only difference is that we need to treat the test and control conditions separately, as we did in the section "Overall Sensitivity and Bias." The detailed calculation is available in supplemental R code and so we will not repeat the calculations here. After the calculation, we have two sets of d' values for each participant: a set of d' values for the test condition and a set of d' values for the control condition. We perform a paired *t*-test to compare these two sets of d' values. This comparison tells us whether our participants' ability to discriminate the unaccusative and unergative stimuli is different in the test condition and the control condition. We found a significant difference between the test and control conditions (t = 9.30, $p < 0.001$). Therefore, our participants differentially discriminated these two types of verbs in these two conditions.

Sensitivity and Bias by Item

It has been argued that, in psycholinguistic research, items should not be treated as a fixed effect (Clark, 1973). It is important to know if the effect we find is driven by certain items or it is true across the board, and therefore it is generally accepted that items should be included as random effects in our statistical models. In this section, we show how to calculate sensitivity and bias in by-items analyses.

In the prenominal participle study, each verb/item appeared in two different conditions: test and control. Each item therefore is associated with four types of responses, as shown in **Table 9**. Here, we want to compare if the response for the test condition is different from that for the control condition. We use the control condition as the baseline because all items in this condition should be acceptable. Therefore an acceptable response in the control condition is a *hit* and an unacceptable response is a *miss*. We assume that an acceptable response in the test condition is a *false alarm* and unacceptable response is a *correct rejection*. With this categorization, if the d' ends up being zero, we know that there is no difference (perceptual distance) between our test and control conditions.

With the above categorization, we can make a frequency table for each item and calculate a d' and a c value for each item. The d' value indicates how different the test condition of the item is from the control condition. The c value indicates if the participants show any response bias for this item.

TABLE 7 | Frequency of the choices in the control condition for the prenominal participle study.

	Unaccusative	Unergative
Acceptable	507	516
Unacceptable	33	24

TABLE 8 | Number of participant responses in each of the four categories defined by the signal detection analysis for the control condition of the prenominal participle study.

Hit (507)	False alarm (516)
Miss (33)	Correct rejections (24)

TABLE 9 | Categorization of judgment data for the prenominal participle study by item.

	Control	Test
Acceptable	Hit	False alarm
Unacceptable	Miss	Correct reject

After calculating the d' for each item, we can assess whether the values for d' in the unaccusative condition are different from those in the unergative condition using a *t*-test. We find a significant difference ($t = -4.37$, $p < 0.005$). However, here we need to be careful with the interpretation of the results. We find that the average d' is larger for the unergative than for the unaccusative condition. Because the d' in our calculation is the perceptual distance between the test condition and the control condition (acceptable condition), the larger this number is, the more different the test condition is from the acceptable condition (less acceptable). Therefore, a larger d' means that the unergative condition is less acceptable. In our example, the larger average d' in the unergative condition means that the unergative condition is less acceptable than the unaccusative condition.

DISCUSSION AND FUTURE DIRECTIONS

In this paper, we first discussed why acceptability judgments can be a useful tool for language research, and we also considered the reliability of the method. Then, we showed how SDT can be applied to analyze the judgment data. After introducing some fundamental concepts, we showed how sensitivity and bias are calculated and how they can help us better interpret acceptability judgment data. In this section, we discuss the assumptions behind the models used in the sections "Signal Detection Theory and One-Factor-Design Experiments" and "Signal Detection Theory and Two-Factor-Design Experiments" and some future directions of research.

The models presented in the sections "Signal Detection Theory and One-Factor-Design Experiments" and "Signal Detection Theory and Two-Factor-Design Experiments" embody two important assumptions: (1) the data follow a Gaussian distribution and (2) the variances of the two distributions are equal. These assumptions are also made by many significant tests such as *t*-test and ANOVA. If the variances are unequal, a single signal detection study will not be sufficient to determine sensitivity and bias. Instead, we will need to have several conditions varying in bias or we will have to conduct a rating-scale experiment (Wickens, 2002; McNicol, 2005). Due to the complexity of this issue, we do not discuss the unequal variance model in this paper. Researchers who are interested in this topic should consult Wickens (2002) and McNicol (2005), among others.

There are some additional interesting questions that can be addressed using SDT. First, it can help us quantify the discriminability of different conditions. Imagine we have three groups of stimuli, Group A (the baseline acceptable control), Group B, and Group C, with stimuli in the two groups differing in their average degree of acceptability. We can calculate a d' using Group A and B which gives us the perceptual distance between Group A and B. We can also calculate a d' using Group A and C which gives us the perceptual distance between Group A and C. Assuming that the d'_{A_B} is 1.2 and d'_{A_C} is 2.2, we can tell that Group B has less perceptual distant from the acceptable condition than Group C (Group B is more acceptable). Although the judgment is binary, d' as a continuous metric can give us a continuous measure of the perceptual distance between different stimuli across a continuum.

We can also compare performance in different populations, which is a more canonical way of using SDT. For example, we can give non-native speakers and native speakers the same stimuli and then compare their performance (d'). If the d' of the native speakers is larger than that of the non-native speaker (as we would expect), we know that native speakers can discriminate the stimuli more accurately, that is, their sensitivity for the phenomenon being tested is better.

There are many remaining questions that need more investigation. In the section "Signal Detection Theory and One-Factor-Design Experiments," we presented one possible measure of bias. We chose this measure to illustrate how bias can be interpreted in the context of acceptability judgments. As we mentioned, there are some alternative measures of bias. Which one best describes the bias in the acceptability judgment data is an empirical question that needs further investigation.

In the paper, we limited our discussion to binary judgments because research has shown that the results for acceptability judgments tend to be consistent regardless of whether the scale provides more than two response categories (Bader and Häussler, 2010). However, we can use SDT for rating judgments involving a non-binary scale as well. One thing to note is that, for acceptability judgments, we usually give participants a scale and ask them to rate the acceptability of the stimuli on that scale. In the context of SDT, rating judgments are performed differently. What participants rate on the scale is not the acceptability of the stimuli but rather how confident they are in their judgment. They still need to make a binary judgment on the acceptability of the stimuli. In addition to that, they need to indicate their confidence level on a scale. One question we can ask is to what degree the acceptability rating and the confidence rating are correlated. Acceptability is believed to be continuous and the gradient judgments from acceptability ratings are believed to reflect the continuous nature of acceptability. However, there is another possibility: the gradient data are created by another factor that is orthogonal to an item's acceptability. One candidate for such an orthogonal factor is confidence level associated with the judgments. By testing the correlation between the acceptability rating and the confidence rating, we can tease apart these two possibilities. If these two factors are uncorrelated, we can exclude the possibility that the gradient judgment is caused by variation in participants' levels of confidence. However, if these two factors correlate significantly, then the gradient data pattern is likely to be caused by participants' confidence level rather than the commonly believed acceptability continuum. In this case, we may need to consider an alternative interpretation of the gradient judgments. It is possible that acceptability is a not a real continuous measure, but the results of these tests are confounded with subjects' confidence about their responses, which is continuous.

SDT can help us address some important questions, including how participants' perceptions of acceptability vary when the linguistic properties of the stimuli are changed in theoretically interesting ways. For example, it is possible to test whether the effect of grammatical violation on acceptability is cumulative. If the effect is cumulative, we would expect stimuli that violate

more rules to be judged less acceptable than stimuli that violate fewer rules. For example, if a set of stimuli violates agreement principles of the grammar whereas another set violates both agreement and case features, the second set should be judged less acceptable than the first set, and this difference should be reflected in their d's. If the ratings of the stimuli can correctly reflect the difference in the degree of acceptability of these stimuli, we expect the d's in these two conditions to differ. We can also change other factors of the stimuli such as the plausibility of the scenario described by the stimuli. This is likely to change participants' judgments: For example, they may judge the more plausible stimuli to be more acceptable. This should happen for both unacceptable and acceptable stimuli. If plausibility and acceptability operate independently, the perceptual distance (d') between these two sets of stimuli should not change because it reflects the acceptability differences between the stimuli. The bias should change because the participants are biased to judge all stimuli to be acceptable. By manipulating different factors in the experiment and seeing how d' and c changes, we can have a better understanding on how plausibility interacts with acceptability. Overall, we believe this approach making use of SDT to analyze binary acceptability responses has the potential to expand our understanding of what such judgments reflect and will allow us to continue to refine our theories of linguistic representation and processing.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the [Open Science Framework] [https://osf.io/pdcye/].

AUTHOR CONTRIBUTIONS

YH and FF conceived of the presented idea, discussed the results, and contributed to the final manuscript. YH developed the theory and performed the computations. FF verified the analytical methods.

FUNDING

This research was partially supported by the National Science Foundation Grant BCS-1650888 to FF.

ACKNOWLEDGMENTS

The authors thank John Henderson and Elaine J. Francis for their discussion of ideas and recommendations of references.

REFERENCES

Achimova, A. (2014). Resolving wh−/quantifier ambiguities: Integrating theoretical and experimental perspectives. Doctoral dissertation. Rutgers University-Graduate School-New Brunswick.

Bader, M., and Häussler, J. (2010). Toward a model of grammaticality judgments. *J. Linguist.* 46, 273–330. doi: 10.1017/S0022226709990260

Borer, H. (1984). "The projection principle and rules of morphology" in *Proceedings of the Fourteenth Annual Meeting of NELS*. eds. C. Jones and P. Sells (Amherst: GLSA, University of Massachusetts), 16–33.

Brysbaert, M., and Stevens, M. (2018). Power analysis and effect size in mixed effects models: a tutorial. *J. Cogn.* 1, 1–20. doi: 10.5334/joc.10

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *J. Verbal Learn. Verbal Behav.* 12, 335–359. doi: 10.1016/S0022-5371(73)80014-3

Clifton, C. Jr., Fanselow, G., and Frazier, L. (2006). Amnestying superiority violations: processing multiple questions. *Linguist. Inquiry* 37, 51–68. doi: 10.1162/002438906775321139

Culicover, P. W., and Jackendoff, R. (2010). Quantitative methods alone are not enough: response to Gibson and Fedorenko. *Trends Cogn. Sci.* 14, 234–235. doi: 10.1016/j.tics.2010.03.012

Edelman, S., and Christiansen, M. H. (2003). How seriously should we take minimalist syntax? A comment on Lasnik. *Trends Cogn. Sci.* 7, 60–61. doi: 10.1016/s1364-6613(02)00045-1

Fabb, N. A. J. (1984). Syntactic affixation. Doctoral dissertation. Cambridge (MA): Massachusetts Institute of Technology.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguist. Rev.* 22, 365–380. doi: 10.1515/tlir.2005.22.2-4.365

Ferreira, F., and Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *J. Mem. Lang.* 30, 725–745. doi: 10.1016/0749-596X(91)90034-H

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013a). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci. USA* 110, 8051–8056. doi: 10.1073/pnas.1216438110

Gibson, E., and Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends Cogn. Sci.* 14, 233–234. doi: 10.1016/j.tics.2010.03.005

Gibson, E., Piantadosi, S. T., and Fedorenko, E. (2013b). Quantitative methods in syntax/semantics research: a response to Sprouse and Almeida. *Lang. Cogn. Process.* 28, 229–240. doi: 10.1080/01690965.2012.704385

Gourevitch, V., and Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika* 32, 25–33. doi: 10.1007/BF02289402

Green, D. M., and Swets, J. A. (1966). *Signal detection theory and psychophysics*. Vol. 1. New York: Wiley.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behav. Res. Methods Instrum. Comput.* 27, 46–51. doi: 10.3758/BF03203619

Huang, Y. (2018). Linking form to meaning: Reevaluating the evidence for the unaccusative hypothesis. Doctoral dissertation. Available at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:40049976

Langsford, S., Perfors, A., Hendrickson, A. T., Kennedy, L. A., and Navarro, D. J. (2018). Quantifying sentence acceptability measures: reliability, bias, and variability. *Glossa: J. Gen. Linguist.* 3, 1–37. doi: 10.5334/gjgl.396

Levin, B., and Rappaport, M. (1986). The formation of adjectival passives. *Linguist. Inquiry* 17, 623–661.

Linzen, T., and Oseki, Y. (2018). The reliability of acceptability judgments across languages. *Glossa: J. Gen. Linguist.* 3:100. doi: 10.5334/gjgl.528

Macmillan, N., and Creelman, C. (2004). *Detection theory: A user's guide*. New York: Psychology Press.

Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychol. Rev.* 84, 452–471. doi: 10.1037/0033-295X.84.5.452

Maling, J., Rizzi, L., and Burzio, L. (1986). *Italian syntax: A government-binding approach. Vol. 1*. Dordrecht: Springer Netherlands.

McNicol, D. (2005). *A primer of signal detection theory*. New York: Psychology Press.

Perlmutter, D. M. (1968). Deep and surface structure constraints in syntax. Doctoral dissertation. MIT.

Perlmutter, D. M. (1978). "Impersonal passives and the unaccusative hypothesis" in *Proceedings of the annual meeting of the Berkeley linguistics society. Vol. 4* (Berkeley: University of California), 157–190. Available at: https://escholarship.org/uc/item/73h0s91v

Peterson, W. W. T. G., Birdsall, T., and Fox, W. (1954). The theory of signal detectability. *Trans. IRE Prof. Group Inf. Theory* 4, 171–212. doi: 10.1109/TIT.1954.1057460

Rugg, M. D., and Curran, T. (2007). Event-related potentials and recognition memory. *Trends Cogn. Sci.* 11, 251–257. doi: 10.1016/j.tics.2007.04.004

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Schütze, C., and Sprouse, J. (2014). "Judgment data" in *Research Methods in Linguistics*. eds. R. Podesva and D. Sharma (Cambridge: Cambridge University Press), 27–50.

Sprouse, J. (2011). A validation of Amazon mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behav. Res. Methods* 43, 155–167. doi: 10.3758/s13428-010-0039-7

Sprouse, J., and Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core syntax. *J. Linguist.* 48, 609–652. doi: 10.1017/S0022226712000011

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua* 134, 219–248. doi: 10.1016/j.lingua.2013.07.002

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instrum. Comput.* 31, 137–149. doi: 10.3758/BF03207704

Swets, J. A., Tanner, W. P. Jr., and Birdsall, T. G. (1961). Decision processes in perception. *Psychol. Rev.* 68, 301–340. doi: 10.1037/h0040547

Tanner, W. P. Jr., and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychol. Rev.* 61, 401–409. doi: 10.1037/h0058700

Wasow, T., and Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua* 115, 1481–1496. doi: 10.1016/j.lingua.2004.07.001

Westfall, J., Kenny, D. A., and Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *J. Exp. Psychol. Gen.* 143, 2020–2045. doi: 10.1037/xge0000014

Wickens, T. D. (2002). *Elementary signal detection theory*. USA: Oxford University Press.

Conflict of Interest: 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

BRIEF RESEARCH REPORT
published: 07 February 2020
doi: 10.3389/fpsyg.2020.00105

Child Relativized Minimality and Grammaticality Judgement

*Anna Gavarró**

Departament de Filologia Catalana, Centre de Lingüística Teòrica, Universitat Autònoma de Barcelona, Barcelona, Spain

Grammaticality judgements are the fundamental experimental source of generative linguistic theory. They may be difficult to elicit, especially in some populations, but generally they inform us neatly about what the grammar licenses or, on the contrary, bans. On the other hand, acceptability is multifactorial and therefore, unlike grammaticality judgement, can be quantified. In this paper I consider a particular empirical domain, that of Relativized Minimality (RM) in acquisition, and its relation to the dichotomy between grammaticality and acceptability. Friedmann et al. (2009) argued that children hold a stricter version of RM than adults. In particular, children would require a disjoint feature specification, not just a distinct feature specification, between target and intervener. The literature shows asymmetries in comprehension of subject and object relative clauses in various languages which fulfill the predictions of child RM. Variation between adults and children might be expected not only in production and comprehension, but also in grammaticality judgement. If so, children would be predicted to reject object relatives as well as the classic RM violations. Alternatively, if child RM is a processing effect, the prediction is that children would be able to tease apart object relative clauses from RM violations under favorable processing conditions. The question I address is: do children assimilate RM violations and object relative clauses? Grammaticality judgement should provide an answer to this question. In this paper I present an experiment targeting grammaticality judgement for object relatives and RM violations and report preliminary results for a group of 6-year-old Catalan-speaking children showing that object relatives and RM violations are not judged in a parallel fashion, since RM violations are rejected more often than object relatives.

Keywords: grammaticality judgement, processing, child grammar, object relatives, Relativized Minimality violations, Catalan

OPEN ACCESS

Edited by:
Urtzi Etxeberria, Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Cristiano Chesi, University Institute of Higher Studies in Pavia, Italy
Evelina Leivada, University of Rovira i Virgili, Spain

***Correspondence:**
Anna Gavarró
anna.gavarro@uab.cat

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

Received: *31 July 2019*
Accepted: *15 January 2020*
Published: *07 February 2020*

Citation:
Gavarró A (2020) Child Relativized Minimality and Grammaticality Judgement. Front. Psychol. 11:105. doi: 10.3389/fpsyg.2020.00105

INTRODUCTION

The literature on language acquisition has attested an asymmetry in the comprehension and production of relative clauses, object relative clauses lagging behind subject relative clauses (see, for English, Brown, 1971; De Villiers et al., 1979; for French, Labelle, 1990; for Portuguese, Corrêa, 1995; for Spanish, Pérez-Leroux, 1995; for German, Adani et al., 2013; for Italian, Contemori and Belletti, 2014, etc.). Friedmann et al. (2009) proposed a new analysis for this well-known asymmetry: subject and object relative clauses differ in the position from which the wh- constituent moves, and they argued that children apply a stricter constraint on A' movement than adults that renders object relatives (under specific conditions) difficult for them. In this report I explore a prediction of Friedmann et al.'s hypothesis if one assumes that this stricter constraint on movement

constitutes a truly grammatical constraint (as opposed to the result of a processing limitation): grammaticality judgement, then, should yield the same pattern as production and comprehension.

The report is organized as follows: in this section I detail Friedmann et al.'s hypothesis and state the prediction to test. In section A Grammaticality Judgement Task, I motivate the experimental design and give details about the experimental items, procedure and participants of the pilot study. In section Results, I present the results, and in section Discussion, I consider them against the background literature.

Following the approach of Relativized Minimality (RM) (Rizzi, 1990) to constraint movement, in a configuration like (1), X and Y fail to relate if Z, the intervener, which is structurally closer to X, can act as its antecedent because of its featural configuration.

(1) … X … Z … Y

The effect of RM can be illustrated with a classic example such as (2), in which movement of *how* is blocked by the intervening interrogative *who*.

(2) *How do you wonder who behaved? ~~how~~
[+Q] [+Q] [+Q]

There is no RM violation in a subject relative clause (3), nor in an object relative clause (4).

(3) the boy that ~~the boy~~ hugs the monkey

(4) the chicken that the cow kisses ~~the chicken~~

Both are well-formed for adults; however, in Friedmann et al.'s (2009) analysis, for children the subject *the cow* acts as an intervener in (4); there is no possible intervener in (3)[1]. This is argued to be the source of children's delay with object relatives. The configuration in (1) can be instantiated as in (5) [(29) in (Friedmann et al., 2009), p. 84].

(5) a.	+A	…	+A	…	<A>	(identity)
b.	+A,+B	…	+A	…	< +A,+B>	(inclusion)
c.	+A	…	+B	…	< +A>	(disjunction)

The example in (2) falls under the case of (5a) and is therefore ill-formed for children and adults alike. When B is featurally distinct from A, as in (5c), the resulting sentence is licensed in both child and adult grammar. Differences only emerge with (5b), where the potential intervener, +A, is characterized by a featural configuration that is a subset of the featural configuration of the antecedent +A+B. This corresponds to the configuration underlying object relatives like (4):

(6) [+R, +NP] … +NP … < +R, +NP>

Adult grammar licenses (6), but child grammar treats it as a violation of (a stronger version of) RM. In Friedmann et al.'s (2009, p. 85) words, "a configuration [like that in (6)] is disallowed as it violates the 'strong' RM requiring featural disjointness." If object relatives[2] are assimilated to RM violations in child grammar, the prediction is then that children will judge them as equally ill-formed in a grammaticality judgement task. This prediction is put to test in the experiment described in the next section.

(5) does not exhaust all possible configurations. In later work on featural RM effects in weak island environments, Villata et al. (2016) consider the configurations in (7).

(7) a.	[+Q]	[+Q]	< +Q>	(bare identity)
b.	[+Q]	[+Q, +N]	< +Q>	(inverse inclusion)
c.	[+Q, +N]	[+Q]	< +Q, +N>	(inclusion)
d.	[+Q, +N]	[+Q, +N]	< +Q, +N>	(complex identity)

In (5) inverse inclusion (7b) was not considered, and bare identity (7a) and complex identity (7d) fell under identity. Inverse inclusion is exemplified in (8a), complex identity in (8b), both of them examples with intervention (taken from Villata et al., 2016, p. 81).

(8) a. Qu'est-ce que$_j$ tu te demandes quel étudiant$_i$ —$_i$ a résolu —$_j$?
what is that you cl-2s wonder which student has solved
'What do you wonder what student solved?'
b. Quel problème$_j$ te demandes-tu quel étudiant$_i$ —$_i$ a résolu —$_j$?
which problem 2s.cl wonder you which student has solved
'Which problem do you wonder which student solved?'

In this paper we focus on the configurations initially considered in Friedmann et al. (2009) and the subsequent research on language acquisition.

A GRAMMATICALITY JUDGEMENT TASK

The experiment designed is a grammaticality judgement task[3]. Young children experience some difficulty in producing grammaticality judgements, possibly because of the inability of the experimenter to transmit what the task is about, and because the task requires some metalinguistic awareness. For that reason, the children recruited were in the age range of 5–7 years and not younger.

[1]Intervention is defined structurally (in terms of c-command), not linearly. This was shown clearly in the case of subject and object relative clauses in Chinese, which follow the asymmetry outlined above in a language in which relative clauses are prenominal and therefore linear and structural intervention do not concur (see Hu, 2014).

[2]To be accurate, not all object relatives are problematic for children (for example, in Hebrew object relatives with null pronominal subjects with arbitrary interpretation, as well as free object relatives, do not give rise to intervention effects), and this follows from different featural specifications, i.e., they would not fall under the configuration (5b)—see Friedmann et al. (2009).

[3]The term I use is grammaticality judgement, as is customary in the generative framework, to refer to the task that a speaker performs when asked about the well-formedness or ill-formedness of a sentence (ill-formedness being represented by an asterisk diacritic); acceptability would refer to well-formedness with respect to a given discourse/pragmatic context, which is not at stake. A sentence is standardly assumed to be grammatical when its derivation converges.

Materials and Methods

Three sentence types were tested in Catalan: (i) object relative clauses, (ii) long distance wh- questions, and (iii) ungrammatical wh- questions involving RM violations. It is worth stressing that the RM violations in the experiment were ill-formed and not just degraded, as some weak island violations may be—see examples (9), taken from Villata et al. (2016) and the examples in Rizzi (1990), as well as the discussion of gradations of acceptability also in relation to RM in Rizzi (2018).

(9) a. ??Which problem do you wonder whether John could solve (in this way)?
b. ?Which problem do you wonder how to solve?

The objective relative in (10) instantiates (5b), the long-distance wh- question in (11) is an instance of (5c), and the wh- question involving a RM violation in (12) instantiates (5a). The featural configuration in (10) is such that the head of the relative clause bears the features [+R,+N], and the intervening DP the feature [+N] (as assumed in Friedmann et al., 2009). The featural configuration of the wh- questions exemplified in (11) is assumed to be [+Q] for the wh- elements involved. Likewise in the ungrammatical question exemplified in (12).

(10) Veig el gos que la nena buscava ~~el gos~~.
see-1s the dog that the girl looked-for
[+R,+N] [+N] <+R,+N>

(11) Com dius que ha vingut ~~com~~?
how say-2s that have-3s come
[+Q] < +Q>

(12) Què penses qui arreglarà ~~què~~?
what think-2s who repair-fut-3s
[+Q] [+Q] < +Q>

In the well-formed wh- questions, two of the experimental sentences contained *dir* "say" as verb selecting the embedded clause and six contained the verb *pensar* "think/wonder"; the same verbs (and in the same proportion) were used in the ungrammatical RM items. The wh- words used were all bare wh- elements, including *què* "what," *qui* "who," *com* "how" and *quan* "when." Since sentences were produced out of context and, furthermore, no complex wh- phrase was used, the effect of D-linking was excluded. In the wh- questions, the wh-element corresponded to an argument or an adjunct of the embedded clause, either because it was a direct object of the embedded verb, or because, as adjuncts, *quan* "when" and *com* "how" would more naturally modify the embedded clause (as in *When do you think you will go?*). The same is true of the RM items: *què* "what" could only be an argument of the embedded clause; in the remaining cases with *com* "how" the adjunct would most naturally modify the embedded verb, *venir* "come over" or *portar-se* "behave," and in this last case it was selected by the verb. In the wh- questions and RM items no overt DP intervened between the wh-elements (subjects were null pronouns except in one case in which the overt subject was postverbal and, therefore, lower in the structure).

Each of the three experimental conditions was exemplified by 6 items, and so the total number of test items was 18 (a complete list appears in the **Annex**). Items were between 7 and 11 syllables long and were presented in pseudorandom order. Of the 18 items, only 6 were ungrammatical for the adult speakers; should children find object relatives ungrammatical, then 12 out of the 18 items would be rejected.

If children assimilate identity configurations (5a) and inclusion configurations (5b), the prediction is that they will perform equally with the two. This is what the literature on child RM has argued: children fail with object relatives when the configuration is that seen in (5b); subject relatives do not give rise to such a configuration, and the subject/object asymmetry follows[4]. A second prediction, not stated by Friedmann et al., is that, if the assimilation of (5b) to (5a) is operative, children will judge instances of (5b) as bad as instances of (5a). This is the rationale of the experiment.

An anonymous reviewer points out that the comparison between object relative clauses and wh- extraction is far from perfect, since these two constructions have been shown to be quite different, so that, for example, in English, Preposition stranding is favored in indirect object wh- questions, but pied-piping is preferred in indirect object relative clauses (Bianchi and Chesi, 2015); in a cross-linguistic study, Sprouse et al. (2016) show that island effects are different between relative clauses and wh- dependencies (in English, adjunct relative clauses do not show island effects, while adjunct wh- dependencies do; Italian does not exhibit subject island effects in relative clauses, but it does in wh- extraction). The reviewer suggests that a better design would therefore include only wh- questions; this remains for future research.

'I see the dog that the girl was looking for.'

'How are you saying he came?'

'What do you think who will repair?'

Participants

The children who participated in the study were native speakers of Central Catalan from the extended metropolitan area of Barcelona. Twenty-five children were tested, but three were excluded because they failed to understand the task. The remaining 22 children were in the age range of 5;05,20 to 7;04,27 (mean age: 6;05). Five adults took part in the experiment as a control group.

The guidelines of the Declaration of Helsinki on human experimentation were enforced during the whole procedure and

[4]The analysis of relative clauses assumed in this literature (and here) is a raising analysis (see Bianchi, 2002a,b).

the experiment was approved by the ethics committee of the UAB (CEEAH evaluation number 4,856).

Procedure

The experiment was carried out individually in a quiet classroom of the children's school. It involved two experimenters, one manipulating a dog puppet and uttering the target sentences, the other introducing the task and questioning the child. The child was told that the puppet was learning to speak but sometimes said things that didn't sound right and so the child would be asked if the sentences s/he heard uttered by the puppet sounded right. The experimental phase was preceded by a training phase consisting of at least two items, one grammatical, another ungrammatical (*Tinc molta gana* "I am very hungry" vs. **Molta tinc gana* "Very I am hungry"); if necessary, the training phase included more items. Positive feedback was given to the child in the experimental phase irrespective of his/her answers. The task took around 15 min. Adults were tested individually on the university campus.

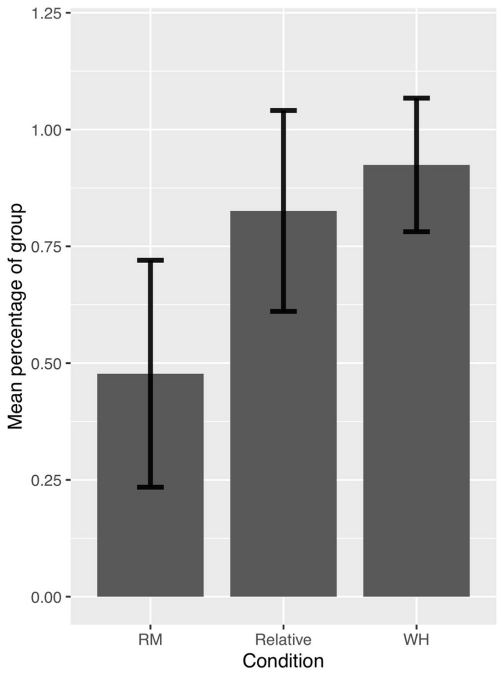


FIGURE 1 | Acceptance rate of the three sentence types, children.

The answers of all participants were recorded on an answer sheet by the second experimenter and then transcribed into an RStudio file.

RESULTS

Adults performed as expected: they rejected all RM violations and accepted all grammatical long- distance interrogatives and object relatives.

The total number of answers provided by the children was 396 (18 × 22), 132 per condition. The data set is freely available at https://ddd.uab.cat/record/215041. Children performed as shown in **Figure 1** and **Table 1**, representing mean acceptance rate, standard deviation and the five number summary (order statistics) Minimum, Q1, Median, Q3, and Maximum.

If we turn to individual results, all the children rejected at least one RM violation, while 10 children accepted all object relative clauses. Two children judged these two sentence types identically; three more children judged RM violations better than object relatives. The remaining 17 children accepted object relatives more often than they accepted RM violations, tending toward the adult pattern. Individual results appear in **Figure 2**.

Even though few children took part in the experiment, and it would be desirable to run it with more participants, some statistical analysis was undertaken. A Generalized Linear Mixed Model was used to model the number of acceptances by sentence type as a binomial response, taking into account repeated measures from each participant. The statistical analysis was obtained using R (R Core Team, 2019).

Statistically significant differences were found as an effect of Sentence Type (F-Value = 63.19; p_value < 0.0001). For the RM (ungrammatical) items, the percentage of estimated acceptance responses was 47.48% ($CI_{95\%}$ = [36.8%, 58.4%]). For the object relative items, the percentage of estimated acceptance responses was 84.22% ($CI_{95\%}$ = [75.59%, 90.19%]). For wh- questions, the percentage of estimated acceptance responses was 93.47% ($CI_{95\%}$ =[87.4%, 96.72%]). These results are represented in **Figure 3**.

Pairwise comparisons of the three sentence types were all significant. There were statistically significant differences between object relatives and RM (z-ratio = 5.8; p_value < 0.0001), with higher acceptance of object relatives than RM violations (OR = 5.9, i.e., the odds ratio of acceptance of object relatives was 5.9 times the odds of acceptance of RM violations). There were marginal statistically significant differences between object relatives and wh- questions (z-ratio = −2.4; p_value = 0.0424), with higher acceptance of wh- questions (OR = 0.37, i.e., the odds ratio of acceptance of wh- questions was 1/0.37≈2.68 times the odds for object relatives). Finally, there were statistically

TABLE 1 | Acceptance of the three sentence types, Mean, SD, and order statistics.

Stype	Data	Mean	SD	Minimum	Q1	Median	Q3	Maximum
Relative	22	0.826	0.215	0.167	0.667	0.833	1.000	1.000
RM	22	0.477	0.243	0.000	0.333	0.417	0.667	0.833
Wh-	22	0.924	0.143	0.000	0.875	1.000	1.000	1.000

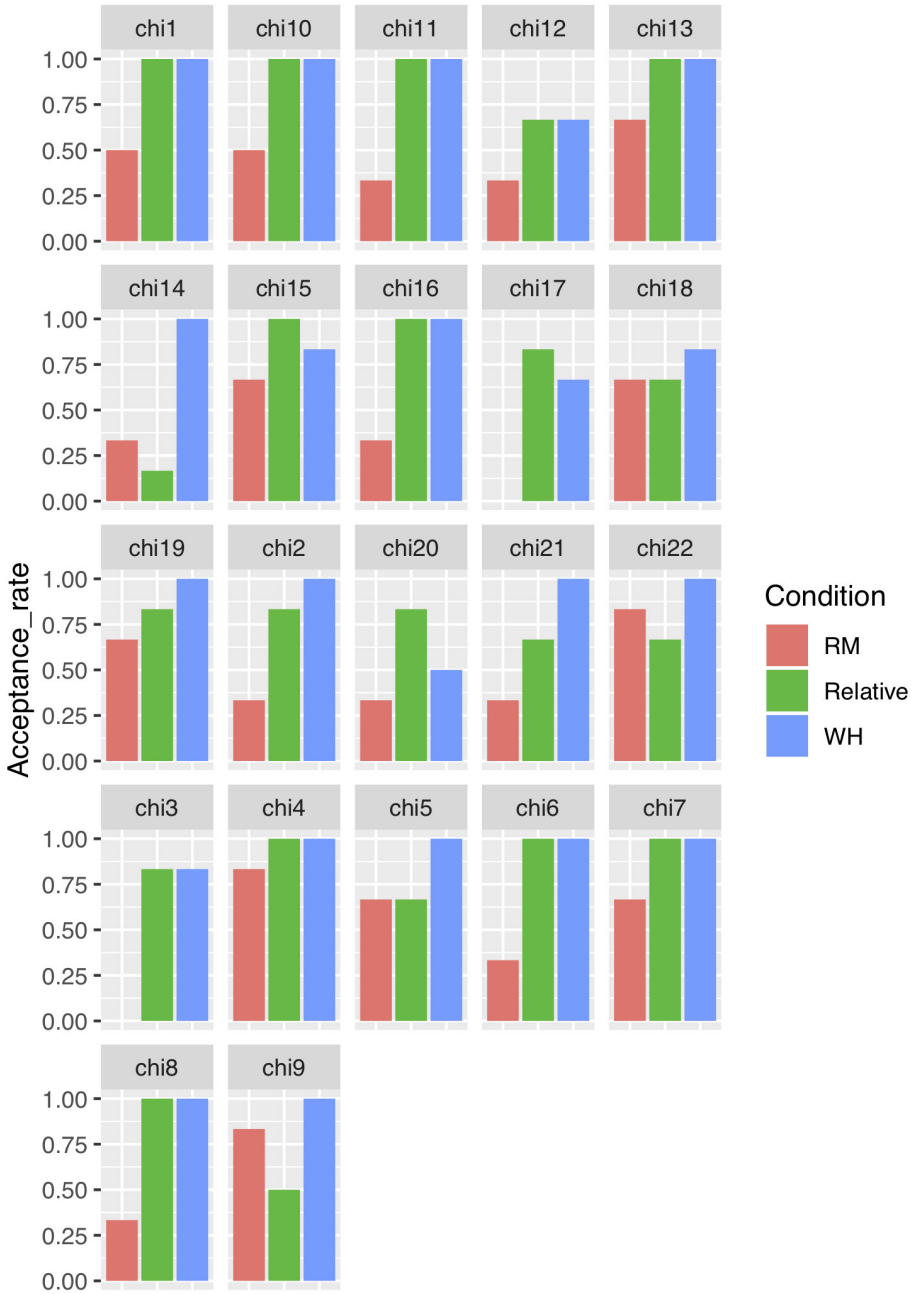


FIGURE 2 | Individual results.

significant differences between RM violations and wh- questions (z-ratio = −7.05; *p*_value < 0.0001), with wh- questions being accepted more often than RM violations (OR = 0.0632, i.e., the odds ratio of acceptance of wh- questions was 1/0.0632≈15.82 times the odds for the RM items).

The results so far show that object relatives and RM violations did not pattern alike for children: children accepted object relatives at much higher rates than RM violations. Rather, object relatives tended to pattern with long-distance wh- questions, as in adult judgements. However, there is a difference in the acceptance rates of object relatives and wh- questions in the judgements of children that is not found in the judgements of adults, albeit the difference is smaller than between any of these two grammatical sentence types and the ungrammatical RM sentences.

These results are tentative; however, with the sample here the hypothesis that object relatives and RM violations are judged in the same way by children cannot be upheld.

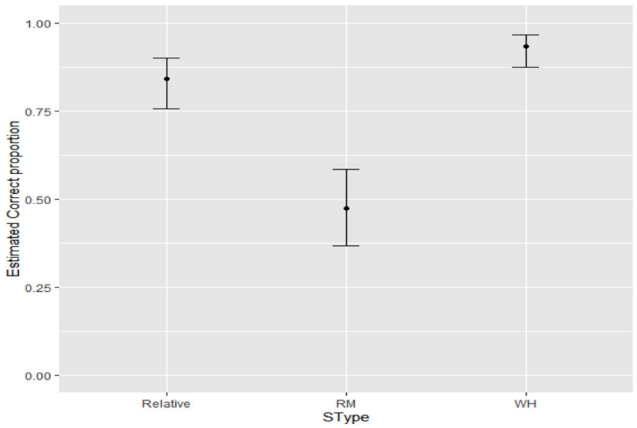


FIGURE 3 | Estimated acceptance of the three sentence types.

TABLE 2 | Subject and object relative clause comprehension and production, Catalan (Gavarró et al., 2012, p. 194).

	Subject relatives		Object relatives[a]	
Comprehension				
4;06–5;06 (Mean 4;11,06)	64/66	97%	53/121	43%
>5;06 (Mean 6;0,12)	60/60	100%	63/110	57%
Total	124/126	98%	116/231	50.2%
Production				
5 (Mean 5;05,15)		98%		62.5%

[a] *The object relatives here include relative clauses with pre- and post-verbal subjects.*

DISCUSSION

In this section, I discuss the results in two respects: first, I consider age and performance in other tasks which, by hypothesis, relate to the one here; second, I go back to the question that motivated this study, namely, does child RM stem from a property of child grammar, defining grammaticality, or does it stem from processing limitations?

First let us consider the results with respect to age. In future research more children and from a wider age range should be tested; with the current sample, the five children who could be considered to conform to the parallel performance in RM violations and object relatives were not amongst the youngest, and performance appears to bear no relation to age (within the limited age span here).

Notice that the children in this study were slightly older than the Hebrew-speaking children in Friedmann et al. (2009), who were in the age range of 3;07 to 5;0. Other studies, however, show that delay in the comprehension of object relatives extends beyond age 5;0. In a study of the acquisition of relative clauses in Catalan, Gavarró et al. (2012), on the basis of a picture identification task, found that the comprehension of object relatives was delayed when compared to the comprehension of subject relative clauses. Production (elicitation based on Novogrodsky and Friedmann's, 2006 method) yielded very similar results [see **Table 2**, which summarizes the results of the two experiments, administered to 21 children (comprehension) and 20 children (production)].

Similar results have been obtained for other languages, such as Italian. In Arosio et al. (2009), which involved 139 Italian-speaking children of ages 5–11, object relatives with post-verbal subjects were miscomprehended at ages 7 and 9 (with adult performance below 50%) and only at 11 was comprehension adult-like (see also Adani, 2010). Parallel results for object wh-interrogatives (also subsumed by Friedmann et al.'s account) showed that 8- to 9-year-olds had not yet achieved adult performance (De Vincenzi et al., 1999; Guasti et al., 2012).

It is beyond the scope of this report to sum up the literature that has been carried out on relative clauses and related constructions over the years, which has led to the development of experiments manipulating Case, number, and gender features (Guasti et al., 2008; Adani et al., 2010; Belletti et al., 2012; Bentea et al., 2016; Friedmann et al., 2017), all relevant to the RM hypothesis[5]. Although Friedmann et al. (2009, p. 71) assert that "the difficulty with object relatives is overcome at around the age of 6 (Friedmann and Novogrodsky, 2004)," the literature on the acquisition of Romance shows that object relatives are not target-like at age 7 (and even beyond) and so, if all of these results are to receive a unified account (a desirable outcome), then we can assume that child RM is operative at age 7, the oldest age group in this study.

To my knowledge, no study so far has considered the child version of RM with grammaticality judgement. The general expectation is that grammaticality judgment should align with production and comprehension, in absence of any indication to the contrary. While dissociations between e.g. production and comprehension are attested in language acquisition, they call for an explanation. If the path of language development is grammar-driven, the prediction is that production, comprehension and judgement will develop in parallel. This is the assumption underlying the experiment in this report. Even though children are known to often fail in their production and comprehension of object relatives in Catalan, and this is attributed to child RM in the literature, they do not judge object relatives in the same way as they judge RM violations. This argues against an assimilation of object relatives and RM violations in child grammar (that is, against the grammatical assimilation of the identity and inclusion conditions).

The results here are exploratory; let us suppose that children do not judge RM violations in the same way they judge object relatives at an age at which the child strict version of RM is operative, as the results so far suggest. In that case, what could the explanation be? Friedmann et al. (2009) do not discard the idea that child RM is the result of a processing limitation. The

[5]There is also work disputing the claims of Friedmann et al. (2009) (see, for example, Goodluck, 2010), and some results that the hypothesis cannot readily encompass, especially from studies on topicalization (e.g., Hu et al., 2018 on Chinese)—but this is not discussed in this paper.

fact that other populations (language impaired children, patients with aphasia) also perform differently with subject and object relatives, and healthy adults under certain circumstances may also display the same asymmetry (Cohen and Mehler, 1996, and much subsequent work; Warren and Gibson, 2002; see Grillo, 2008) would seem to favor a processing account. In Friedmann et al.'s words (2009, p. 84–85), "It may be tempting to speculate that the ban against [(5b)] in early systems may relate to a limitation in the operative syntactic memory: clearly, disjointness is easier to determine, as it can be calculated feature by feature, whereas calculating a subset-superset relation requires holding in operative memory and comparing the whole featural specifications associated with different positions, an operation that may exceed the capacity of the early systems." In adults, on the other hand, "a partial overlap of features giving rise to a configuration like [(5b)] is grammatical, but determines 'complexity effects' detectable in experimental work." Under such a processing account, one could speculate that the source of the difference between the results here and the results in the literature on relative clause comprehension and production are related to the experimental method. If grammaticality judgement is less costly than comprehension/production in terms of processing (to the extent that the interpretation of the sentence may not need to be fully accessed) then one would predict that object relatives and RM violations would not be judged homogeneously by children, even under the assumption that child RM is operative.

In addition, there is a further difference between object relatives and RM violations, even for children: while children do produce (to varying degrees) object relatives, RM violations of the kind exemplified in (2) and (12) are not attested. This may be an indication that the configuration underlying object relatives is part of child grammar, while RM violations are ungrammatical for children. Grammaticality judgement can therefore provide a new source of evidence to characterize child RM as either a grammatical or a processing phenomenon.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are freely available at: https://ddd.uab.cat/record/215041.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Comissió d'Ètica en l'Experimentació Animal i Humana (CEEAH), Universitat Autònoma de Barcelona. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AG designed the experiment reported and ran it with the collaboration of Míriam Muntané. AG analyzed the data with the help of Ester Boixadera and wrote the paper.

FUNDING

The author acknowledges the following funding sources: 2017 SGR 634 from the Generalitat de Catalunya, and project FFI2017-87699-P from the Ministerio de Economía y Competitividad.

ACKNOWLEDGMENTS

As well as the funding agencies mentioned above, the author wishes to thank Míriam Muntané for acting as experimenter; the Escola Maria Borés in La Pobla de Claramunt, for allowing the participation of their children in the study; and the children who took part in the experiment for their enthusiastic collaboration. The author thanks the two reviewers and J.M. Brucart for their suggestions, and Ester Boixadera, from the Servei d'Estadística Aplicada of the UAB, for conducting the statistical analysis; any remaining errors are the author's sole responsibility.

REFERENCES

Adani, F. (2010). Rethinking the acquisition of relative clauses in Italian: towards a grammaticality based account. *J. Child Lang.* 38, 141–165. doi: 10.1017/S0305000909990250

Adani, F., Sehm, M., and Zukowski, A. (2013). "How do German children and adults deal with their relatives," in *Advance of Language Acquisition*, eds S. Stavrakaki, M. Lalioti, and P. Konstantinopolou (Newcastle: Cambridge Scholars Publishing), 14–22.

Adani, F., Van Der Lely, H. K. J., Forgiarini, M., and Guasti, M. T. (2010). Grammatical feature dissimilarities make relative clauses easier: a comprehension study with Italian children. *Lingua* 120, 2148–2166. doi: 10.1016/j.lingua.2010.03.018

Arosio, F., Adani, F., and Guasti, M. T. (2009). "Grammatical features in the comprehension of Italian relative clauses by children," in *Merging Features: Computation, Interpretation, and Acquisition*, eds M. José Brucart, A. Gavarró, and J.Solà (Oxford; New York, NY: Oxford University Press), 138–155.

Belletti, A., Friedmann, N., Brunato, D., and Rizzi, L. (2012). Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua* 122, 1053–1069. doi: 10.1016/j.lingua.2012.02.007

Bentea, A., Durrleman, S., and Rizzi, L. (2016). Refining intervention: the acquisition of featural relations in object A-bar dependencies. *Lingua* 169, 21–41. doi: 10.1016/j.lingua.2015.10.001

Bianchi, V. (2002a). Headed relative clauses in generative syntax. Part I. *Glot. Int.* 6, 197–204.

Bianchi, V. (2002b). Headed relative clauses in generative syntax. Part II. *Glot Int.* 6, 1–13.

Bianchi, V., and Chesi, C. (2015). "On a PP/DP asymmetry in extraction," in *Structures, Strategies and Beyond. Studies in Honour of Adriana Belletti,* eds E. Di Domenico, C. Hamann, and S. Matteini (Amsterdam; Philadelphia, PA: John Benjamins), 47–66. doi: 10.1075/la.223.03bia

Brown, H. D. (1971). Children's comprehension of relativized English sentences. *Child Dev.* 42, 1923–1936.

Cohen, L., and Mehler, J. (1996). Click monitoring revisited: an on-line study of sentence comprehension. *Mem. Cogn.* 24, 94–102.

Contemori, C., and Belletti, A. (2014). Relatives and passive object relatives in Italian-speaking children and adults: intervention in production and comprehension. *Appl. Psychol.* 35, 1021–1053. doi: 10.1017/S0142716412000689

Corrêa, L. (1995). An alternative assessment of children's comprehension of relative clauses. *J. Psychol. Res.* 24, 183–203. doi: 10.1007/BF02145355

De Villiers, J., Tager Flusberg, H. B., Hakuta, K., and Cohen, M. (1979). Children's comprehension of relative clauses. *J. Psychol. Res.* 17, 57–64.

De Vincenzi, M., Arduino, L. S., Ciccarelli, L., and Job, R. (1999). "Parsing strategies in children comprehension of interrogative sentences," in *Proceedings of European Conference on Cognitive Science,* ed S. Bagnara (Rome: Istituto di Psicologia del CNR), 301–308.

Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives: types of intervention in the acquisition of A-bar dependencies. *Lingua* 119, 67–88. doi: 10.1016/j.lingua.2008.09.002

Friedmann, N., and Novogrodsky, R. (2004). The acquisition of relative clause comprehension in Hebrew: a study of SLI and normal development. *J. Child Lang.* 31, 661–681. doi: 10.1017/S0305000904006269

Friedmann, N., Rizzi, L., and Belletti, A. (2017). No case for case in locality: case does not help interpretation when intervention blocks A-bar chains. *Glossa J. Gen. Linguist.* 2:33. doi: 10.5334/gjgl.165

Gavarró, A., Cunill, A., Muntané, M., and Reguant, M. (2012). The acquisition of catalan relatives: structure and processing. *Revue Roumaine de Linguist.* 57, 183–201.

Goodluck, H. (2010). Object extraction is not subject to child relativized minimality. *Lingua* 120, 1516–1521. doi: 10.1016/j.lingua.2009.10.005

Grillo, N. (2008). *Generalized minimality: syntactic underspecification in Broca's*. (Doctoral dissertation). LOT, Universiteit Utrecht, Utrecht.

Guasti, M. T., Branchini, C., and Arosio, F. (2012). Interference in the production of Italian subject and object wh-questions. *Appl. Psycholinguist.* 33, 185–223. doi: 10.1017/S0142716411000324

Guasti, M. T., Stavrakaki, S., and Arosio, F. (2008). "Number and case in the comprehension of relative clauses: evidence from Italian and Greek," in *Language Acquisition and Development. Proceedings of GALA 2007*, eds A. Gavarró and M. J. Freitas (Newcastle: Cambridge Scholars Publishing), 230–240.

Hu, S. (2014). *Intervention effects and the acquisition of relativization and topicalization in Chinese*. (Doctoral dissertation). Universitat Autònoma de Barcelona and Università degli Studi di Milano-Bicocca, Milan.

Hu, S., Guasti, M. T., and Gavarró, A. (2018). Chinese children's knowledge of topicalization: experimental evidence from a comprehension study. *J. Psycholinguist. Res.* 47, 1279–1300. doi: 10.1007/s10936-018-9575-6

Labelle, M. (1990). Predication, wh-movement and the development of relative clauses. *Lang. Acquisit.* 1, 95–119.

Novogrodsky, R., and Friedmann, N. (2006). The production of relative clauses in syntactic SLI: a window to the nature of the impairment. *Adv. Speech-Lang. Pathol.* 8, 364–375. doi: 10.1080/14417040600919496

Pérez-Leroux, A.-T. (1995). Resumptives in the acquisition of relative clauses. *Lang. Acquisit.* 4, 105–138. doi: 10.1080/10489223.1995.9671661

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rizzi, L. (1990). *Relativized Minimality*. Cambridge, MA: The MIT Press.

Rizzi, L. (2018). Intervention effects in grammar and language acquisition. *Probus* 30, 339–367. doi: 10.1515/probus-2018-0006

Sprouse, J., Caponigro, I., Greco, C., and Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Nat. Lang. Linguist. Theory* 34, 307–344. doi: 10.1007/s11049-015-9286-8

Villata, S., Rizzi, L., and Franck, J. (2016). Intervention effects and relativized minimality. new experimental evidence from graded judgments. *Lingua* 179, 76–96. doi: 10.1016/j.lingua.2016.03.004

Warren, T., and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition* 85, 79–112. doi: 10.1016/s0010-0277(02)00087-2

Conflict of Interest: 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.

ANNEX: EXPERIMENTAL ITEMS

Object relatives within transitive clauses (1–6), RM violations (7–12), long-distance interrogatives (13–18):

1.– Veig la nena que la mestra ha renyat.
2.– Veig el gos que la nena buscava.
3.– Tinc el pinzell que els nens buscaven.
4.– Veig el gat que el gos ha mossegat.
5.– Veig els pollets que la gallina buscava.
6.– Tinc el conte que els nens llegiran.
7.– Com dius qui es porta?
8.– Com dius qui vindrà?
9.– Com penses qui vindrà?
10.– Què penses com farà?
11.– Què penses qui arreglarà?
12.– Què penses qui llegirà?
13.– Com dius que ha vingut?
14.– Com dius que va a casa seva?
15.– Què penses que farem avui?
16.– Què penses que farà la senyoreta?
17.– Quan penses que farem vacances?
18.– Quan penses que aniràs a casa?

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 February 2020
doi: 10.3389/fpsyg.2020.00186

Processing Prescriptively Incorrect Comparative Particles: Evidence From Sentence-Matching and Eye-Tracking

Ferdy Hubers[1†], Theresa Redl[1,2†], Hugo de Vos[3], Lukas Reinarz[4] and Helen de Hoop[1]*

[1] Centre for Language Studies, Radboud University Nijmegen, Nijmegen, Netherlands, [2] Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands, [3] Institute of Public Administration, Leiden University, Leiden, Netherlands, [4] Department of Physics, University of Bonn, Bonn, Germany

Speakers of a language sometimes use particular constructions which violate prescriptive grammar rules. Despite their prescriptive ungrammaticality, they can occur rather frequently. One such example is the comparative construction in Dutch and similarly in German, where the *equative* particle is used in comparative constructions instead of the prescriptively correct *comparative* particle (Dutch *beter als Jan* and German *besser wie Jan* "lit. better as John"). In a series of three experiments using sentence-matching and eye-tracking methodology, we investigated whether this grammatical norm violation is processed as grammatical, as ungrammatical, or whether it falls in between these two. We hypothesized that the latter would be the case. We analyzed our data using linear mixed effects models in order to capture possible individual differences. The results of the sentence-matching experiments, which were conducted in both Dutch and German, showed that the grammatical norm violation patterns with ungrammatical sentences in both languages. Our hypothesis was therefore not borne out. However, using the more sensitive eye-tracking method on Dutch speakers only, we found that the ungrammatical alternative leads to higher reading times than the grammatical norm violation. We also found significant individual variation regarding this very effect. We furthermore replicated the processing difference between the grammatical norm violation and the prescriptively correct variant. In summary, we conclude that while the results of the more sensitive eye-tracking experiment suggest that the grammatical norm violation is not processed completely on a par with ungrammatical sentences, the results of all three experiments clearly show that the grammatical norm violation cannot be considered grammatical, either.

Keywords: grammatical norm violations, comparative particles, sentence-matching, eye-tracking, grammaticality

OPEN ACCESS

Edited by:
M. Teresa Espinal, Autonomous University of Barcelona, Spain

Reviewed by:
Gosse Bouma, University of Groningen, Netherlands
Katharina Spalek, Humboldt University of Berlin, Germany

***Correspondence:**
Theresa Redl
Theresa.Redl@mpi.nl

[†] These authors have contributed equally to this work

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

Received: *17 September 2019*
Accepted: *27 January 2020*
Published: *14 February 2020*

Citation:
Hubers F, Redl T, de Vos H, Reinarz L and de Hoop H (2020) Processing Prescriptively Incorrect Comparative Particles: Evidence From Sentence-Matching and Eye-Tracking. Front. Psychol. 11:186. doi: 10.3389/fpsyg.2020.00186

INTRODUCTION

Decades of experimental research contrasting grammatical with ungrammatical sentences have taught us much about language processing (e.g., Hagoort et al., 1993; Friederici et al., 2006). But what exactly constitutes an ungrammatical sentence? To a linguist, a grammatical sentence is one that adheres to the natural rules and constraints of a native speaker's grammar, produced and

understood by those exposed to the same input; ungrammatical sentences are constructions that are in principle not generated by a native speaker's competence, although grammaticality judgments may vary (Schütze, 1996). To many language users, in contrast, an ungrammatical sentence is one that is prescriptively "incorrect" and is not, or rather in their view should not, be part of the standard language. These definitions clash when considering constructions that are frequently produced and encountered by native speakers of the language, yet nevertheless firmly disapproved of by speakers who adhere to prescriptive grammar rules. Consider the following sentence in (1), a well-known example of such a construction in Dutch (Hubers and de Hoop, 2013).

(1) *Jane is sterker als Jackie.*
Jane is stronger as Jackie
"Jane is stronger than (lit. *as*) Jackie."

In (1) the use of *als* "as" instead of *dan* "than" is deemed incorrect by speakers of Dutch, even though prescriptive language guides often acknowledge the existence of this variant in a comparative construction (van der Meulen, 2018). Comparatives in Dutch (and German) belong to the category of "linguistic constructions where at least some degree of unwanted variation exists or is thought to exist" (van der Meulen, 2018: 79). Prescriptively disapproved variants such as *als* "as" in (1) are called *grammatical norm violations* (Hubers et al., 2016) or *grammatical taboos* (Vogel, 2019).

It remains unclear how such grammatical norm violations are processed by the human brain. Do they pattern with grammatical sentences, do they pattern with ungrammatical sentences, or do they fall in between grammatical and ungrammatical constructions when it comes to processing? The present paper employs two experimental techniques to investigate the online processing of grammatical norm violations compared to the processing of both grammatical and ungrammatical sentences. Before we continue to discuss these experiments in Sections "Experiments 1 and 2: Sentence-Matching" and "Experiment 3: Eye-Tracking," Section "Grammaticality vs. Acceptability" reviews literature on (the processing of) ungrammaticality vs. unacceptability, focusing on one particular grammatical norm violation in Dutch (and German), namely, the use of an equative particle in a comparative construction as illustrated in (1) above.

GRAMMATICALITY VS. ACCEPTABILITY

In an acceptability judgment study, Vogel (2019) found that grammatical taboos are judged as marked, but not to the same degree as truly ungrammatical sentences. He carried out three different types of acceptability tasks, one asking for aesthetic judgments, one for normative judgments (i.e., whether the construction was considered prescriptively correct), and one for possibility (i.e., estimated occurrence frequency). The grammatical taboos were judged in between grammatical (unmarked) filler sentences and ungrammatical ones, and behaved approximately on a par with the linguistically marked filler sentences in all three types of judgment experiments. Both types of markedness received a similar mean value in the different acceptability tasks, although grammatical taboos were disapproved of more strongly than linguistically marked sentences under the aesthetic judgment test. The two most salient grammatical taboos even grouped together with the ungrammatical sentences under the aesthetic judgment test. Yet, what Vogel considered the strongest grammatical taboo in his study, that is, the use of the verb *tun* "do" as an auxiliary in German, overall received a significantly higher acceptability rate than the ungrammatical filler sentences. In fact, it even came out as grammatical under the possibility type of judgment test.

Vogel (2019) also raised the question whether his empirical method could distinguish between grammatical taboos and linguistically marked sentences, based on their source of markedness. The markedness of linguistically marked sentences has a grammar-internal cause, whereas the markedness of grammatical taboos is caused externally, namely, by a social norm and about ten years of education. Vogel (2019) indeed found a difference between the two types of markedness (internally or externally caused), namely, in the between-subject variance. Participants were more uniform in their judgments of linguistically marked sentences than in their judgments of grammatical taboos. Clearly, grammatical taboos are not always marked or unacceptable for everybody, as there are many speakers who actually use these constructions themselves, perhaps even unaware of their low sociolinguistic status in prescriptive grammar.

Focusing on language users who are definitely aware of this lower prestige, Hubers et al. (2016) set up an fMRI study in which they presented sentences containing grammatical norm violations as well as grammatical and ungrammatical sentences. Their participants were recruited on the basis of their knowledge of prescriptive grammar rules, but also because of their strong negative attitude toward grammatical norm violations. To test whether social cognition was involved in the processing of grammatical norm violations, Hubers et al. also compared grammatical norm violations to sentences describing violations of social norms in their experiment. The latter type of sentences did not contain a linguistic violation, hence were grammatical. The authors did not find any effects specific to the processing of grammatical norm violations, whereas they did so for social norm violations. Also, during the processing of grammatical norm violations, some brain regions were activated that were also involved in the processing of ungrammatical sentences. No evidence for overlapping brain regions was found for social norm violations in comparison with ungrammatical sentences. These results suggest that grammatical norm violations are purely linguistic violations and not social ones. Still, grammatical norm violations are not completely ungrammatical, since Hubers et al. (2016) also found that similar brain regions were involved in both the processing of grammatical sentences and grammatical norm violations. Their explanation was that both types of sentences can be interpreted and integrated with conceptual memory equally well.

The present paper further investigates the online processing of grammatical norm violations compared to the processing of both grammatical and ungrammatical sentences using two

other experimental techniques than Hubers et al. (2016). The first method we apply is a sentence-matching task, first used by Forster (1979). In a sentence-matching task, participants are sequentially presented with two sentences, and they have to indicate whether the second sentence is identical to the first one or not. Identical grammatical sentences are confirmed faster than identical ungrammatical sentences (Forster, 1979; Freedman and Forster, 1985; Forster and Stevenson, 1987; Duffield et al., 2002, 2007). Duffield et al. (2007) use the task to investigate a French construction that is deemed ungrammatical, but that is processed in the same way as grammatical sentences. Duffield et al. conclude that the construction is underlyingly grammatical. Hubers et al. (2016), on the other hand, found an increased activation in Inferior Frontal Gyrus for grammatical norm violations, similar to the activation found for ungrammatical sentences as opposed to grammatical sentences (Hagoort, 2005; Friederici et al., 2006; Snijders et al., 2009), but they also found processing overlap between grammatical norm violations and truly grammatical sentences. On the basis of the elicited reaction times, the sentence-matching task provides a straightforward method to find out whether grammatical norm violations as in (1) above are processed as either grammatical or ungrammatical, or indeed somewhere in between. The latter result could be concluded from reaction times slower than those for ungrammatical sentences but faster than those for grammatical and prescriptively correct ones. This would indicate an intermediate grammaticality status, in accordance with the results of Hubers et al. (2016) and Vogel (2019). The second method we apply is eye-tracking. This paradigm can give more information on the time course of the processing of the grammatical norm violation. If the ungrammaticality judgment of a grammatical norm violation is the result of a more conscious process than in the case of truly ungrammatical sentences, then we may expect the processing difficulties that arise with grammatical norm violations to be more global and to occur only later in the reading process. Therefore, tracking the eye-movements of participants incrementally reading these grammatical norm violations in comparison with reading grammatical and ungrammatical sentences will provide a valuable addition to the overall reaction time data from the sentence-matching task.

In the current paper, the grammatical norm violation under study concerns the use of an equative particle in a comparative construction. We focus on this particular grammatical norm violation, because it is one of the few violations that are prominent in both Dutch and German. Moreover, we did not include other grammatical norm violations in our study, since these were not expected to lead to different results, as was also shown in a *post hoc* analysis by Hubers et al. (2016). No processing differences were found between the various grammatical norm violations included in their study.

Before we continue to discuss these two types of experiments in Sections "Experiments 1 and 2: Sentence-Matching" and "Experiment 3: Eye-Tracking", the remainder of Section "Grammaticality vs. Acceptability" reviews the particular grammatical norm violation in Dutch (and German) that constitutes the focus of the present paper, the use of an equative particle in a comparative construction, as illustrated in (1) above.

Grammaticality vs. Acceptability in Comparative Particles in Dutch and German

Recall example sentence in (1) above, repeated below for convenience, which reflects the Dutch transition from the use of the comparative particle *dan* "than" toward the equative particle *als* "as" in comparatives (Reinarz et al., 2016).

(1) *Jane is sterker als Jackie.*
Jane is stronger as Jackie
"Jane is stronger than (lit. *as*) Jackie."

In present-day German a similar process takes place, in which the equative particle *wie* "as" is used in comparatives, instead of the comparative particle *als* "than" (Jäger, 2010). Whereas different theoretical linguistic analyses have been proposed to account for this replacement of a comparative particle by an equative particle in comparatives (cf. Postma, 2006; Reinarz et al., 2016; Jäger, 2019), the underlying idea of all of these theoretical analyses is that there must be an important grammar-internal factor in the grammars of Dutch and German inducing it. Postma (2006) argues that this factor is that the particle *dan* "than" has lost its original negative meaning. Reinarz et al. (2016) argue that the replacement results from a conflict between two competing functional principles, Economy and Iconicity. Jäger (2019) proposes a syntactic reanalysis of comparison constructions as embedded clauses, on the basis of a historically underlying correlative construction. All of these analyses thus assume the change to be internal to the language system, taking place irrespective of external (counter)forces. However, as Milroy and Milroy (1985: 348) state, "some innovations may not be accepted by a community and hence may not lead to change." This adequately characterizes the current state of affairs concerning the use of an equative particle in a comparative in Dutch and German. Hubers and de Hoop (2013) find a strong correlation between level of education and the use of *als* "as" or *dan* "than" in a comparative. They argue that this correlation clearly reflects the strong influence of the prescriptive rule taught in schools (see also Hubers et al., 2019), repressing the use of an equative particle in a comparative construction in Dutch. The prescriptive rule against the use of an equative particle in a comparative construction is a well-known issue in German, too (Grebe, 1966; Jäger, 2010). To sum up, on the one hand, the use of an equative particle in a comparative can be considered a linguistic innovation, which is somehow caused language-internally. On the other hand, the use of a comparative particle in a comparative reflects a language-external prescriptive rule, as prescriptivists are notoriously intolerant of innovations in language. The result of these two counterforces, one language-internal and one language-external, is the extant variation up until now between two particles in comparatives in both Dutch and German.

Whatever motivates the use of an equative particle in a comparative construction in Dutch and German, the fact that it frequently occurs in the language makes this grammatical norm violation different from ordinary ungrammatical sentences, which hardly ever show up in everyday speech. Not only is the

prescriptive rule explicitly taught in secondary education (Hubers et al., 2019), lay people also regularly express their concerns about the grammatical norm violation on social media (cf. Ermans, 2016 on German). The following anecdote from January 2014 serves to illustrate. In a radio interview, the former chair of the Dutch parliament, Anouchka van Miltenburg, said that each day when she got up, she decided to do her job better *als gisteren* "as yesterday." This grammatical norm violation gave rise to so many negative reactions from the audience, especially on Twitter, that when the interview was broadcasted again the next day, one could hear van Miltenburg all of a sudden say that she would do her job better *dan gisteren* "than yesterday." The recorded audio material of the interview had been edited, and van Miltenburg's *als* "as" had been cut out and replaced by *dan* "than."[1]

On the basis of Hubers et al. (2016), we expect to find processing differences between the grammatical norm violations (Dutch *beter als* and German *besser wie* "better as") and the ungrammatical constructions (Dutch *beter wie* and German *besser wer,* "better who"), as well as between grammatical norm violations and their grammatical and prescriptively correct counterparts (Dutch *beter dan* and German *besser als* "better than"). More specifically, we expect that the processing cost due to the grammatical norm violation is smaller and less immediate than the processing cost caused by an ungrammatical construction. This is because the latter is caused by a language-internal type of ungrammaticality, whereas the processing cost linked to the grammatical norm violation is caused externally, by a prescriptive norm, and thus may emerge only after the subject's conscious evaluation of the construction. We will also explore individual differences in the processing of the three constructions.

EXPERIMENTS 1 AND 2: SENTENCE-MATCHING

We conducted two versions of the sentence-matching experiment in order to investigate the processing of grammatical norm violations. Experiment 1 was conducted in Dutch with native speakers of Dutch and Experiment 2 was conducted in German with native speakers of German.

Materials and Methods

Participants

In total, 38 university students participated in Experiment 1 (seven males). They were all native speakers of Dutch, and most of them (35) were right-handed. Participants were recruited through SONA, the participant database of Radboud University. They all had normal or corrected-to-normal vision. The experiment took about 30 min and participants were rewarded with a 5 Euro gift card or course credit.

A total of 92 German university students participated in Experiment 2 (24 males). The data were collected at Radboud University in the Netherlands, University of Cologne, and Free University Berlin in Germany. The majority of the participants were right-handed (81) and all of them had normal or corrected-to-normal vision. Just like Experiment 1, Experiment 2 took about 30 min. Participation was compensated with a 5 Euro gift card.

[1]The news item on this can be found online at: http://www.rtlnieuws.nl/editienl/laatste-videos-editienl/wnl-corrigeert-taalfoutje-van-miltenburg.

Materials

The materials in Experiment 1 consisted of 18 experimental sentences and 76 filler sentences. The experimental sentences all included a comparative construction, and were taken from Hubers et al. (2016). Three versions of the same experimental sentence were created by changing the comparative particle. In the first version, the equative particle *als* "as" was used, resulting in a grammatical norm violation [condition GN, see (2a)]. The second version contained the grammatical and prescriptively correct particle *dan* "than" [condition GC, presented in (2b)], and the third version contained the question word *wie* "who," leading to a truly ungrammatical sentence [condition UG, see (2c)]. The experimental sentences were all presented in the matching condition, requiring a yes response.

(2) a. *Gijs is slimmer als de andere leraren.* (condition GN)
Gijs is smarter as the other teachers
"Gijs is smarter than (lit. *as*) the other teachers."
b. *Gijs is slimmer dan de andere leraren.* (condition GC)
Gijs is smarter than the other teachers
"Gijs is smarter than the other teachers."
c. *Gijs is slimmer wie de andere leraren.* (condition UG)
Gijs is smarter who the other teachers
"*Gijs is smarter who the other teachers."

As reported in Hubers et al. (2016), the experimental sentences were all pretested to see whether they elicited the intended effect. A grammaticality judgment task revealed that more than 80% of the 136 participants judged the grammatical sentences as being grammatically correct, while less than 20% of the participants judged the ungrammatical sentences as being grammatically correct.

The fillers were all grammatical sentences. Based on Duffield et al. (2002), about 40% of the fillers were presented in the non-matching condition, leading to one-third of all materials requiring a no-response. To this end, we created slightly adapted versions of the sentences by changing only one word. An example of a filler sentence pair in the non-matching condition can be found in (3).

(3) a. *De postbode heeft de brief verkeerd bezorgd.*
the mail.carrier has the letter wrong delivered
"The mail carrier wrongly delivered the letter."
b. *De postbode heeft een brief verkeerd bezorgd.*
the mail.carrier has a letter wrong delivered
"The mail carrier wrongly delivered a letter."

In order to control for sentence length, both the filler and experimental sentences consisted of 12 or 13 syllables.

In Experiment 2, the same sentences were used as in Experiment 1, but translated into German by a German-Dutch bilingual student. See (4a–c) for the German equivalents of the Dutch experimental sentences presented in (2). Dutch proper names were replaced by German ones.

(4) a. *Uwe* *ist* *klüger* *wie* *die*
Uwe is smarter as the
anderen *Lehrer.* (condition GN)
other teachers
"Uwe is smarter than (lit. *as*) the other teachers."

b. *Uwe ist* *klüger* *als* *die*
Uwe is smarter than the
anderen *Lehrer.* (condition GC)
ther teachers
"Uwe is smarter than the other teachers."

c. *Uwe* *ist* *klüger* *wer* *die*
Uwe is smarter who the
anderen *Lehrer.* (condition UG)
other teachers
"*Uwe is smarter who the other teachers."

Fillers were also translated into German, see (5) for the translation of example (3).

(5) a. *Der* *Postbote* *hat* *den* *Brief*
the mail.carrier has the letter
falsch *zugestellt.*
wrong delivered
"The mail carrier wrongly delivered the letter."

b. *Der* *Postbote* *hat* *einen* *Brief*
the mail.carrier has a letter
falsch *zugestellt.*
wrong delivered
"The mail carrier wrongly delivered a letter."

Sentence length of the filler and experimental sentences in Experiment 2 were comparable to those of Experiment 1.

Procedure

The procedure of Experiments 1 and 2 was identical. Participants were tested in a sound-attenuated booth. The experiment was conducted using E-prime (Schneider et al., 2002), and a buttonbox was used to record the participants' responses.

A trial started with a fixation cross that was presented at the center of the screen for 250 ms. Subsequently, the first sentence was displayed at the top left of the screen for 3000 ms. Next, the second sentence was presented at the bottom of the screen. Upon the presentation of the second sentence, participants were instructed to indicate as quickly as possible whether the second sentence was identical to the first one by using the buttonbox. The button corresponding with "yes" was always located at the dominant hand. The second sentence disappeared after an answer was given by the participant or after 3000 ms.

The experiment started with a practice session consisting of six practice trials in order for the participant to get used to the task. After the practice session, they had the opportunity to ask questions if anything was unclear. Subsequently, participants completed two blocks of 47 trials separated by a short break. After the experiment, they filled in a background questionnaire. In addition, participants were presented with 10 sentences which they had to correct if necessary. The aim of this test was to assess participants' familiarity with Dutch prescriptive grammar rules.

Analysis

We analyzed the data of Experiments 1 and 2 using linear mixed effects models in R with the lme4 package (Bates et al., 2015). The reaction time data were log transformed to correct for a right skew in the data. The three-level factor *condition* (*als/wie* "as," *dan/als* "than," *wie/wer* "who") was coded using simple contrasts (UCLA and Statistical Consulting Group, 2011). With simple contrasts, the reference level is always coded as −1/3, and the level that is contrasted is coded as 2/3. They are similar to treatment contrasts, but have the advantage that the intercept corresponds to the mean of means instead of corresponding to the mean of the reference level. One contrast compared *als/wie* "as" to *dan/als* "than" (i.e., contrast 1, coded as −1/3, 2/3, −1/3), the other compared *als/wie* "as" to *wie/wer* "who" (i.e., contrast 2, coded as −1/3, −1/3, 2/3). We included random intercepts for items and participants, as well as random slopes for condition for both items and participants. Following Bates et al. (2015), overparameterization was checked by means of principal component analysis. In case of overparameterization, correlation parameters were dropped as a first step. If overparameterization persisted, individual variance components were removed. However, this later step was never necessary, and all models were fitted with the full random structure excluding correlation parameters. *P*-values were obtained using the package lmerTest (Kuznetsova et al., 2017). Finally, in order to gauge the individual variation between participants, we carried out two different model comparisons. We compared the full model to a model without random slopes for contrast 1 per participant, and we also compared the full model to a model without random slopes for contrast 2 per participant. This was done to see whether the two variance components significantly increased the model fit. In other words, we tested whether individual variation was significant.

Results

Experiment 1

The analyses were conducted on the correct responses on the experimental sentences only. To this end, 5.5% of the data was removed. The averaged logarithmically transformed reaction times and standard deviations of the experimental sentences per condition are visualized in **Figure 1**. The mean reaction times and SDs on the response scale (in milliseconds) are presented in **Supplementary Table S1**.

The linear mixed effects regression analysis revealed a significant effect of *als* vs. *dan* (beta = −0.09, SE = 0.022, $t = -3.87$, $p < 0.01$). Participants took longer to decide that the sentences were identical if the sentences contained

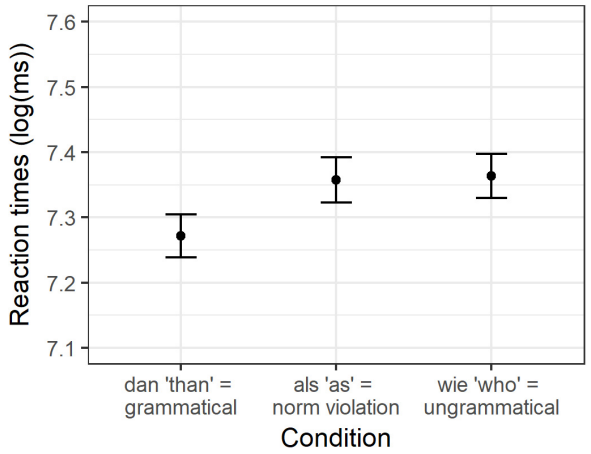


FIGURE 1 | Condition means for logarithmically transformed reaction times in Experiment 1.

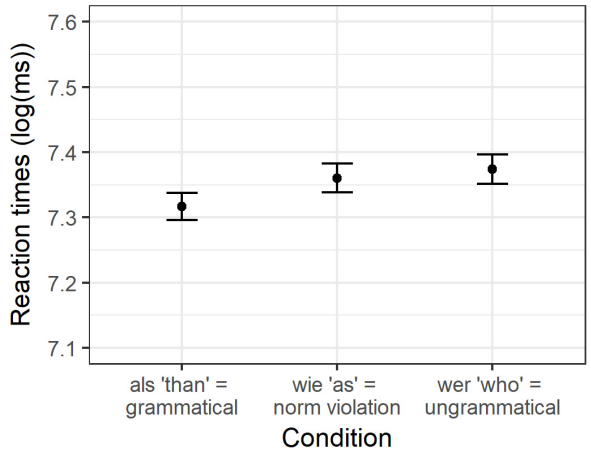


FIGURE 2 | Condition means for logarithmically transformed reaction times in Experiment 2.

the grammatical norm violation (*als* "as") as compared to its grammatically correct counterpart (*dan* "than"). Decisions to sentence pairs containing *als* "as" did not significantly differ from decisions to sentences containing *wie* "who" (beta = 0.01, SE = 0.024, $t = 0.318$, $p = 0.75$).

Gauging individual differences

As can be seen from **Table 1**, adding random slopes per participant for the contrast comparing *als* and *wie* did not significantly improve the full model. The same result was found for random slopes per participant for the contrast comparing *als* and *dan*. This suggests no significant individual variation among participants.

Experiment 2

As in Experiment 1, only the experimental sentences were included in the analyses that were responded to correctly. Therefore, 5.7% of the data had to be removed. **Figure 2** displays the average logarithmically transformed reaction times and the corresponding SDs for the experimental items per condition. The mean reaction times and SDs on the response scale are presented in **Supplementary Table S2**.

The analysis showed a significant effect between *wie* "as" and *als* "than" (beta = −0.04, SE = 0.019, $t = -2.181$, $p < 0.05$). More specifically, German participants were slower to decide that a sentence pair was identical if it contained a grammatical norm violation (the particle *wie* "as") than if it was grammatical. No significant difference was observed between sentences containing a grammatical norm violation (*wie* "as") and the ungrammatical *wer* "who" (beta = 0.02, SE = 0.015, $t = 1.127$, $p = 0.27$).

Gauging individual differences

Similar to Experiment 1, the model did not significantly improve after adding random slopes per participant for the contrast comparing *wie* "as" and *als* "than" and random slopes per participant for the contrast comparing *wie* "as" and *wer* "who" (see **Table 2**).

Discussion

Unlike what we had predicted on the basis of Hubers et al. (2016), we did not find a difference between the processing of the grammatical norm violation *beter als/besser wie* "better as" and the ungrammatical condition *beter wie/besser wer* "better who," whereas the processing of the grammatical norm violation differed significantly from the processing of its grammatical and prescriptively correct counterpart *beter dan/besser als* "better than." Apparently, when participants have to determine whether two sentences are identical, the grammatical norm violation slows down this process to the same extent as the ungrammatical sentence. If it is true that a sentence-matching task provides us with a better measure of grammaticality than a grammatical judgment task, as claimed by Duffield et al. (2002, 2007), we must conclude that at least this particular grammatical norm violation is not underlyingly grammatical, but plainly ungrammatical.

However, in a sentence-matching task, the processing of a sentence is measured only after the full sentence has been read and can be judged. We assume that our participants, who were mostly university students, are well aware of the grammatical norm violation when they encounter it. Yet, it may be that this realization does not arise immediately. That is, if the grammatical

TABLE 1 | Results of significance tests for random slope components per participant in Experiment 1.

	Chi squared (*df* = 1)	*p*-value
Contrast 1 (als vs. dan)	0	1
Contrast 2 (als vs. wie)	0.132	0.717

TABLE 2 | Results of significance tests for random slope components per participant in Experiment 2.

	Chi squared (*df* = 1)	*p*-value
Contrast 1 (wie vs. als)	0.016	0.9
Contrast 2 (wie vs. wer)	1.597	0.206

norm violation is in fact underlyingly grammatical, processing will not be hampered immediately, but rather only when the linguistic awareness has arisen that the sentence is a grammatical norm violation. In order to find out whether processing a grammatical norm violation is comparable to processing a grammatical sentence in its initial stage of processing, we decided to conduct an eye-tracking experiment. We hypothesize that grammatical sentences will be processed with most ease, and that grammatical norm violations will possibly pattern with them. Ungrammatical sentences will lead to the largest processing cost.

EXPERIMENT 3: EYE-TRACKING

Materials and Methods

Participants

The data were collected as part of another experiment, for which our stimuli functioned as fillers. Due to the design of this other experiment, participants were not equally distributed over our three lists, leading to, respectively, 45, 45, and 30 participants. We excluded 22 participants that grew up with the Limburgian dialect, or had a background in linguistics. We excluded Limburgian participants because in their dialect *wie* "who" as a particle in comparative constructions does occur. Participants with a background in linguistics were excluded because they might be aware of the phenomenon of grammatical norm violations. After excluding these participants, we were left with data of 36 participants for lists 1 and 2, and data of 26 participants for list 3. Subsequently, we randomly excluded 10 participants from lists 1 and 2 to match the number of participants in list 3. This led to a total of 78 participants (31 male) to be included in the analysis. These participants answered at least 75% of correction questions pertaining to filler items correctly. The participants ranged in age from 18 to 28 (M = 21.8) and were all native speakers of Dutch. Participants were recruited through SONA, the participant database of Radboud University. All participants had normal or corrected-to-normal vision. The experiment took approximately one hour, and reimbursement was a 10 Euro coupon or course credit, if preferred.

Materials

Each participant saw 18 stimuli in three conditions. These stimuli were based on the experimental sentences included in the Dutch version of the sentence-matching experiment (Experiment 1). Each item occurred in every condition, and three lists were created to be equally distributed across participants.

Procedure

The experiment was conducted at the Centre for Language Studies labs at Radboud University. We used an EyeLink 1000 + remote desktop eye-tracker with a chinrest to stabilize the participants' head. Viewing was binocular and the participants' dominant eye was sampled at 1000 Hz. If it was not possible to sample the dominant eye (e.g., due to glare which often occurs when participants wear glasses), the other eye was measured. This was the case for 14 participants. The stimuli were presented at a distance of 108 cm on a BenQ XL 2420T 24" LED using Experiment Builder by SR Research. The stimuli were presented in 19-point Calibri font on a light gray background. Upon arrival, participants were given an information document about the experiment and asked to sign a consent form. The experimenter then determined the participant's dominant eye. They were then accompanied to the testing booth where they read the instructions. The experimenter then set up the eye-tracker and made sure participants were comfortable. Participants then performed a 13-point calibration and validation. They then saw four practice items, after which they got the opportunity to ask questions. After another calibration routine, the experiment started. Participants got the opportunity to take breaks after one- and two-thirds of the experiment. After the experiment, participants filled in a short questionnaire, probing participants for the purpose of the experiment. Finally, they were paid.

Analysis

The raw eye-tracking data were pre-processed using EyeLink Data Viewer by SR Research. Using this software, we examined the fixation pattern of each item for each participant. The fixations were reassigned in case a clear shift had occurred. Furthermore, we deleted the first fixation of a trial if it did not fall on the first line of the stimulus. Fixations that were smaller than 80 ms were merged with another fixation within 0.25 degrees (i.e., within 0.47 cm) in visual angle on the *x*-axis if this fixation exceeded 80ms. Subsequently, unmerged fixations that were larger than 1200 ms or smaller than 80 ms were deleted. We calculated three reading time measures for the regions of interest: first run dwell time (i.e., the sum of the duration of all fixations in a region when it is entered for the first time), regression path duration (i.e., first run dwell time with the addition of the duration of fixations back to previous regions out of the analyzed region), and dwell time (i.e., the sum of the duration of all fixations in a region, also known as total fixation duration). The two regions of interest were defined as follows, as indicated by the square brackets:

(6) *De* *koffie* *is* *inderdaad* *sterker*
the coffee is indeed stronger
[*dan/als/wie*] [*gisteren*]
than/as/who yesterday
maar *ik* *vind* *hem* *zo* *wel* *lekker*
but I find him so PRT nice
en *bovendien* *kun* *je*
and moreover can you
er *nog* *melk* *door* *doen* *als* *je* *wilt*
there still milk through do as you want
"The coffee is indeed stronger [than/as/who] [yesterday], but I like it better like this and you can still add milk if you want."

The comparative particle was always followed by the word *gisteren* "yesterday" in order to keep the spillover region constant. Both the particle and the spillover region were analyzed.

TABLE 3 | Condition means and SDs in milliseconds on particle and spillover region for first run dwell time, regression path duration, and dwell time.

	Reading time measure					
	First run dwell time		Regression path duration		Dwell time	
	M	*SD*	*M*	*SD*	*M*	*SD*
Particle						
dan	244	127	370	310	308	182
als	259	143	392	329	427	343
wie	266	135	474	405	606	501
Spillover						
dan	235	126	359	336	312	196
als	260	141	509	531	439	328
wie	296	196	748	725	675	546

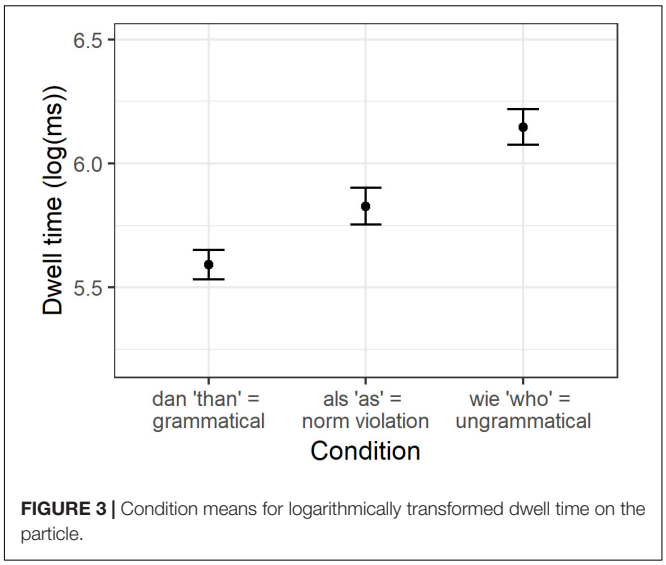


FIGURE 3 | Condition means for logarithmically transformed dwell time on the particle.

We analyzed the data using linear mixed effects models in R with the lme4 package (Bates et al., 2015). The three different reading time measures were log transformed to correct for a right skew in the data. The data were analyzed in the same way as in the sentence-matching experiments (see section "Analysis").

Results

The condition means are shown in **Table 3**. No significant effects were found for first run dwell time. For regression path duration, we found a significant increase in reading time for *wie* compared to *als* (beta = 0.17, SE = 0.06, $t = 2.940$, $p = 0.006$), but not for *als* compared to *dan* (beta = −0.04, SE = 0.05, $t = -0.875$, $p = 0.38$). For dwell time, however, we found that both *dan* (beta = −0.22, SE = 0.05, $t = -4.311$, $p < 0.001$) and *wie* (beta = 0.34, SE = 0.05, $t = 7.359$, $p < 0.001$) differed significantly from *als* (see **Figure 3**). For the spillover region, we found that first run dwell time was significantly higher after *als* compared to *dan* (beta = −0.09, SE = 0.03, $t = -2.504$, $p = 0.02$), and after *wie* compared to *als* (beta = 0.09, SE = 0.03, $t = 2.715$, $p = 0.008$). The results for regression path duration were again rather similar, with an increase after *als* compared to *dan* (beta = −0.27, SE = 0.05, $t = -5.34$, $p < 0.001$) and an increase after *wie* compared to *als* (beta = 0.33, SE = 0.07, $t = 4.761$, $p < 0.001$). And finally, we found the same results again for dwell time on the spillover region: *als* led to an increase compared to *dan* (beta = −0.28, SE = 0.04, $t = -6.543$, $p < 0.001$) and *wie* led to an increase compared to *als* (beta = 0.36, SE = 0.04, $t = 8.074$, $p < 0.001$).

To sum up, the earliest effect was found on the comparative particle for regression path duration: ungrammatical *wie* took significantly longer than the norm violation *als*, but we found no difference for *als* compared to grammatical *dan*. For dwell time on the particle, we found that both *dan* and *wie* differed significantly from *als*. This finding persisted for the spillover region for all three reading times: *als* leads to an increase in reading time compared to *dan*, but *wie* slows reading down even more.

Gauging Individual Differences

As can be seen in **Table 4**, we found that random slopes per participant for the contrast comparing *als* and *wie* significantly improved the model for regression path duration on the particle itself, as well as for first run dwell time and regression path duration in the spillover region. Thus, for earlier reading times we found significant individual variation regarding the effect of *als* compared to *wie*. Note that we did not find significant individual variation for first run dwell time on the particle, but this is not surprising, given that we found no significant effects whatsoever for this model. This suggests that participants showed larger individual variation regarding the comparison between *als* and *wie*, while no such individual differences were found for the contrast comparing *als* and *dan*. **Figure 4** shows this individual variation for first run dwell time on the spillover region.

Discussion

By employing the eye-tracking method, we hoped to gain more insight into the immediate processing of grammatical norm violations compared to a grammatical and ungrammatical alternative, as well as potential changes over time. The earliest significant effect which surfaced suggests that ungrammatical *wie* "who" leads to an increase in processing effort compared to the grammatical norm violation *als* "as." Interestingly, the difference between grammatical *dan* "than" and the grammatical norm violation *als* "as" was not yet significant in this model. Later on in processing, however, we consistently found that the grammatical norm violation leads to higher reading times than its grammatical counterpart, while the ungrammatical variant leads to higher reading times than the grammatical norm violation. The grammatical norm violation seems to fall in between the grammatical and the ungrammatical alternative. Furthermore, we found that our participants showed large individual variation regarding the comparison between the grammatical norm violation and the ungrammatical alternative, but not for the comparison between the grammatical option and the grammatical norm violation. In other words, while on a group level, ungrammatical *wie* "who" led to higher reading times than the grammatical norm violation *als* "as", the size of this effect differed vastly between participants. For

TABLE 4 | Results of significance tests for random slope components per participant.

	Reading time measure					
	First run dwell time		Regression path duration		Dwell time	
	Chi squared (*df* = 1)	*p*-value	Chi squared (*df* = 1)	*p*-value	Chi squared (*df* = 1)	*p*-value
Particle						
Contrast 1 (als vs. dan)	0	1	0	1	0	1
Contrast 2 (als vs. wie)	1.883	0.170	4.911	0.027*	0	1
Spillover						
Contrast 1 (als vs. dan)	0.426	0.514	0.132	0.717	1	0
Contrast 2 (als vs. wie)	4.983	0.026*	5.304	0.021*	1.498	0.221

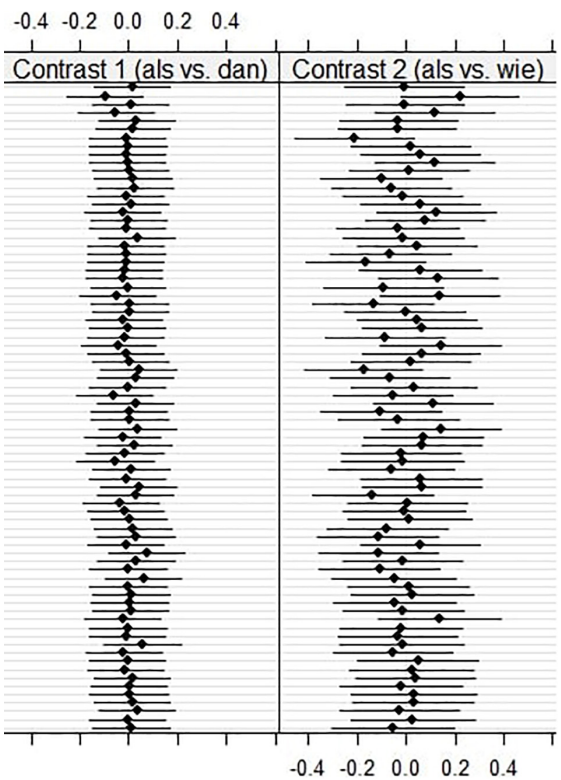


FIGURE 4 | Plots for conditional modes and 95% prediction intervals of all 78 participants for random slopes of both tested contrasts for first run dwell time on the spillover region (see Kliegl et al., 2011).

some participants, *als* "as" was in fact just as bad as *wie* "who". The effect that the grammatical norm violation led to increased processing compared to the grammatical alternative was consistent across participants.

The earliest effect was found on the particle itself, and in this initial stage we did not find any difference between the grammatical particle *dan* "than" and the prescriptively incorrect particle *als* "as", whereas the ungrammatical particle *wie* "who" gave rise to an increase in processing cost. One might wonder whether in this initial stage, the sentence is simply not yet ungrammatical, since adding a comparative phrase is not obligatory, and *als* "as" could also be used to introduce a conditional adjunct clause, as in (7).

(7) *De koffie is inderdaad sterker als*
the coffee is indeed stronger as
je er geen melk
you there no milk
door doet.
through does
"The coffee is indeed stronger if you don't add milk."

Notoriously, however, people do not read words in an isolated fashion, but are able to preview the upcoming characters and word(s) (e.g., McConkie and Rayner, 1975). Thus, our participants are likely to already have parsed the grammatical norm violation *sterker als gisteren* "lit. stronger as yesterday" as it was intended when we measure their reading time on *als* "as". This could be seen as weak evidence that participants do indeed process the grammatical norm violation as grammatical and not as ungrammatical in the initial stage of processing. Later on, their processing of the grammatical norm violation falls in between the processing of grammatical and ungrammatical alternatives. This is in accordance with Hubers et al.'s (2016) findings, who assume that processing grammatical norm violations is partly like processing grammatical sentences because both are perfectly interpretable, and partly like processing ungrammatical sentences because they violate a grammatical rule, irrespective of whether the source of the rule is grammar-internal or grammar-external (prescriptive).

GENERAL DISCUSSION

This study investigated the processing of a salient grammatical norm violation in Dutch and German, whether native speakers process such grammatical norm violations as underlyingly grammatical, as ungrammatical, or as somewhere in between. In a series of acceptability judgment tests, Vogel (2019) found that students of German judged grammatical norm violations as equally marked as linguistically marked expressions, that is, in between grammatical and ungrammatical sentences. However, participants showed less uniformity in their judgments

of grammatical norm violations than in those of linguistically marked expressions. Hubers et al. (2016) conducted an fMRI study and concluded that grammatical norm violations were processed differently from both grammatical and ungrammatical sentences. The participants in their study were between 30 and 50 years old, and especially selected because they strongly disapproved of grammatical norm violations. It could be that younger speakers who may not hold such a very strong view on grammatical norm violations or who even use these constructions themselves, process them as underlyingly grammatical. This could also be predicted on the basis of Duffield et al. (2002, 2007) who found that certain constructions in French, which were judged ungrammatical in grammaticality judgment tasks, were processed like grammatical sentences in a sentence-matching experiment, which led them to conclude that they were underlyingly grammatical. The authors concluded from this that a sentence-matching task may be a more reliable tool than a traditional grammaticality judgment task in revealing the grammaticality of a construction.

Two versions of the same sentence-matching experiment, a Dutch and a German one, were carried out with university students in the Netherlands and Germany. We hypothesized that the participants would process the grammatical norm violation under consideration as in between grammatical and ungrammatical, as expected on the basis of Hubers et al.'s (2016) fMRI study, and in line with Vogel's (2019) acceptability judgment tests, but not in accordance with Duffield et al. (2002, 2007) findings that certain constructions that are judged ungrammatical are processed like grammatical ones. Strikingly, however, the results of the sentence-matching experiments showed no difference at all between the processing of grammatical norm violations and ungrammatical sentences, whereas there was a clear difference between the grammatical norm violations and the grammatical sentences. Hence, for the Dutch and German university students in our experiment, processing a grammatical norm violation is just as problematic as processing an ungrammatical sentence. There was no indication whatsoever that grammatical norm violations could be considered underlyingly grammatical.

Various linguistic case studies have shown the influence of prescriptive grammar rules on language use and language change unambiguously (e.g., Hubers and de Hoop, 2013; Hinrichs et al., 2015). Yet, linguists generally consider grammatical norm violations to be grammatical, simply because they frequently occur in the language under consideration, which means they can be generated and understood by the grammatical system. By contrast, the majority of the speakers of a language, in particular educated speakers such as the university students used in our experiments, may be convinced that grammatical norm violations are ungrammatical, as this is what they have learned in school or at home. Vogel (2019: 48) calls this a "paradox": a grammatical norm violation is generated by the principles of the grammar, otherwise it would not occur at all in the language, but because it violates socially induced grammatical norms, speakers believe that the construction cannot (or should not) be part of that grammar. The linguistic awareness of prescriptive grammar rules may account for the fact that the participants processed grammatical norm violations in the same way as ungrammatical sentences in the sentence-matching experiment. Because a sentence-matching task is a purely linguistic task, participants are very much focused on the grammatical form of the sentence. Also, the decision that has to be made in a sentence-matching experiment, namely, whether the second sentence is identical to the first or not, requires careful consideration of the entire sentence. In the final stage of processing, when they have to make the decision, participants will generally be aware of the presence of a grammatical norm violation, which they have read even twice in a row (because the two experimental sentences were always identical in the task). Supposedly, this explains their delay in reaction time. Note that this type of processing does not reflect what is going on in everyday speech, where grammatical norm violations may often remain unnoticed, because they occur rather frequently, and because they are perfectly interpretable.

In order to find out whether grammatical norm violations are being processed as grammatical or ungrammatical in a somewhat more natural type of setting, we conducted an eye-tracking experiment. In this experiment, sentences were only presented once, and there was no additional linguistic task. Besides, unlike in the sentence-matching experiment, the processing was measured incrementally, that is, right from the beginning of the reading process. Here we found an immediate difference between the processing of ungrammatical sentences and grammatical norm violations. In the very first stage of processing, we did not find a difference between the processing of grammatical norm violations and that of grammatical sentences, while ungrammatical sentences did already lead to an increase in processing cost. After this initial stage of processing, the grammatical norm violation in the eye-tracking study consequently behaved in between grammatical and ungrammatical sentences, confirming the findings of Hubers et al. (2016) as well as Vogel (2019). Also, we found more variation among the participants regarding the difference between the grammatical norm violation and the ungrammatical alternative. No significant individual differences were found for the comparison between the grammatical norm violation and the grammatical alternative. The first encounter with the grammatical norm violation did not lead to a significant increase in processing cost, unlike ungrammaticality. However, after this initial stage processing difficulties do occur, although overall less severe than in the case of the ungrammatical alternative. For some participants, however, the grammatical norm violation is just as bad as the ungrammatical sentence, as reflected in individual differences.

CONCLUSION

The aim of this paper was to shed light on the processing of grammatical norm violations or grammatical taboos (Hubers et al., 2016; Vogel, 2019). We focused on one grammatical norm

violation in particular, namely, the use of an equative particle in a comparative construction, which occurs frequently in Dutch as well as German, and which is well-known for being a violation of prescriptive grammar rules. We investigated whether this grammatical norm violation gets processed as grammatical or ungrammatical or as something in between.

The results of two sentence-matching experiments, one in Dutch and one in German, indicated that the grammatical norm violation was processed as ungrammatical. However, we hypothesized that this might be explained by the fact that in a sentence-matching task processing is only measured after the full sentence has been taken into account, at which point participants are probably fully aware of the grammatical norm violation they have encountered.

Evidence for this hypothesis was obtained from an eye-tracking experiment, in which the difference between the grammatical norm violation and the ungrammatical alternative immediately surfaced, while we did not find a difference in processing between the grammatical norm violation and its grammatical counterpart early on. Later on in processing, the grammatical norm violation consistently fell in between the grammatical and ungrammatical variants. Moreover, the difference between the grammatical norm violation and the ungrammatical alternative showed a large amount of individual variation, suggesting that for some language users the grammatical norm violation was just as bad as the ungrammatical alternative.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was reviewed and approved by the Ethics Assessment Committee (EAC) of the Faculty of Arts and the Faculty of Philosophy, Theology and Religious Studies of Radboud University Nijmegen (numbers 6889 and 4592), in line with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study, manuscript revision, and read and approved the submitted version. FH, TR, and HV were involved in the data collection. FH, TR, and HH wrote the manuscript. FH and TR performed the statistical analyses.

ACKNOWLEDGMENTS

The Dutch sentence-matching experiment was carried out within the Radboud Honours Academy Program, which is gratefully acknowledged. We would like to thank Marieke Ermans and Johanna Longerich for collecting the data of the German sentence-matching experiment. We would further like to thank the two reviewers, as well as our colleagues Thijs Trompenaars and Marten van der Meulen for feedback on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00186/full#supplementary-material

REFERENCES

Bates, D. M., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48.

Duffield, N., Matsuo, A., and Roberts, L. (2007). Acceptable ungrammaticality in sentence matching. *Second Lang. Res.* 23, 155–177. doi: 10.1177/0267658307076544

Duffield, N., White, L., De Garavito, J. B., Montrul, S., and Prévost, P. (2002). Clitic placement in L2 French evidence from sentence matching. *J. Linguist.* 38, 487–525. doi: 10.1017/s0022226702001688

Ermans, M. (2016). *Besser wie als. The Acceptance of wie as a Comparative Particle in German*. MA thesis, Radboud University, Nijmegen.

Forster, K. (1979). "Levels of processing and the structure of the language processor," in *Sentence Processing*, eds W. E. Cooper, and E. C. T. Walker, (Hillsdale, NJ: Lawrence Erlbaum).

Forster, K. I., and Stevenson, B. J. (1987). Sentence matching and well-formedness. *Cognition* 26, 171–186. doi: 10.1016/0010-0277(87)90029-1

Freedman, S. E., and Forster, K. I. (1985). The psychological status of overgenerated sentences. *Cognition* 19, 101–131. doi: 10.1016/0010-0277(85)90015-0

Friederici, A. D., Fiebach, C. J., Schlesewsky, M., Bornkessel, I. D., and von Cramon, D. Y. (2006). Processing linguistic complexity and grammaticality in the left frontal cortex. *Cereb. Cortex* 16, 1709–1717. doi: 10.1093/cercor/bhj106

Grebe, P. (1966). Sprachnorm und Sprachwirklichkeit. *Wirkendes Wort* 16, 145–156.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

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

Hinrichs, L., Szmrecsanyi, B., and Bohmann, A. (2015). Which-hunting and the Standard English relative clause. *Language* 91, 806–836. doi: 10.1353/lan.2015.0062

Hubers, F., and de Hoop, H. (2013). The effect of prescriptivism on comparative markers in spoken Dutch. *Linguist. Netherlands* 2013, 89–101. doi: 10.1075/avt.30.07hub

Hubers, F., Snijders, T. M., and de Hoop, H. (2016). How the brain processes violations of the grammatical norm: an fMRI study. *Brain Lang.* 163, 22–31. doi: 10.1016/j.bandl.2016.08.006

Hubers, F., Trompenaars, T., Collin, S., de Schepper, K., and de Hoop, H. (2019). Hypercorrection as a by-product of education. *Appl. Linguist.* amz001. doi: 10.1093/applin/amz001

Jäger, A. (2010). Der Komparativzyklus und die Position der Vergleichspartikeln. *Linguistische Berichte* 224, 467–493.

Jäger, A. (2019). The syntax of comparison constructions in diachronic and dialectal perspective. *Glossa* 4:70.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., and Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention. *Front. Psychol.* 1:238. doi: 10.3389/fpsyg.2010.00238

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

McConkie, G. W., and Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Percept. Psychophys.* 17, 578–586. doi: 10.1016/j.visres.2017.06.005

Milroy, J., and Milroy, L. (1985). Linguistic change, social network and speaker innovation. *J. Linguist.* 21, 339–384. doi: 10.1017/s0022226700010306

Postma, G. (2006). Van *groter dan* naar *groter als* — structurele oorzaken voor het verval van het comparatieve voegwoord *dan*. *Nederlandse Taalkunde* 11, 2–22.

Reinarz, L., de Vos, H., and de Hoop, H. (2016). Conflicting constraints in the comparative cycle. *J. German Linguist.* 28, 403–425. doi: 10.1017/S1470542716000131

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime (Version 2.0). [Computer Software and Manual]*. Pittsburgh, PA: Psychology Software Tools Inc.

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago, IL: The University of Chicago Press.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J. A., Petersson, K. M., and Hagoort, P. (2009). Retrieval and unification of syntactic structure in sentence comprehension: an FMRI study using word-category ambiguity. *Cereb. Cortex* 19, 1493–1503. doi: 10.1093/cercor/bhn187

UCLA and Statistical Consulting Group (2011). *R Library Contrast Coding Systems for Categorical Variables*. Available at: https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/#SIMPLE (accessed July 29, 2019).

van der Meulen, M. (2018). Do we want more or less variation? The comparative markers *als* and and *dan* in Dutch prescriptivism since 1900. *Linguist. Netherlands* 2018, 79–96. doi: 10.1075/avt.00006.meu

Vogel, R. (2019). Grammatical taboos. An investigation on the impact of prescription in acceptability judgement experiments. *Zeitschrift für Sprachwissenschaft* 38, 37–79. doi: 10.1515/zfs-2019-0002

Conflict of Interest: 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

HYPOTHESIS AND THEORY
published: 10 March 2020
doi: 10.3389/fpsyg.2020.00364

Acceptable Ungrammatical Sentences, Unacceptable Grammatical Sentences, and the Role of the Cognitive Parser

Evelina Leivada[1] and Marit Westergaard[2,3]*

[1] Universitat Rovira i Virgili, Tarragona, Spain, [2] Arctic University of Norway, Tromsø, Norway, [3] Norwegian University of Science and Technology, Trondheim, Norway

A search for the terms "acceptability judgment tasks" and "language" and "grammaticality judgment tasks" and "language" produces results which report findings that are based on the exact same elicitation technique. Although certain scholars have argued that acceptability and grammaticality are two separable notions that refer to different concepts, there are contexts in which the two terms are used interchangeably. The present work reaffirms that these two notions and their scales do not coincide: there are sentences that are acceptable, even though they are ungrammatical, and sentences that are unacceptable, despite being grammatical. First, we adduce a number of examples for both cases, including grammatical illusions, violations of Identity Avoidance, and sentences that involve a level of processing complexity that overloads the cognitive parser and tricks it into (un)acceptability. We then discuss whether the acceptability of grammatically ill-formed sentences entails that we assign a meaning to them. Last, it is shown that there are *n* ways of unacceptability, and two ways of ungrammaticality, in the absolute and the relative sense. Since the use of the terms "acceptable" and "grammatical" is often found in experiments that constitute the core of the evidential base of linguistics, disentangling their various uses is likely to aid the field reach a better level of terminological clarity.

Keywords: grammaticality, grammatical illusions, syntactic islands, parser, processing

OPEN ACCESS

Edited by:
Urtzi Etxeberria, Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Roelien Bastiaanse, University of Groningen, Netherlands
Kepa Erdocia, University of the Basque Country, Spain

****Correspondence:***
Evelina Leivada
evelina@biolinguistics.eu

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

Received: *23 August 2019*
Accepted: *17 February 2020*
Published: *10 March 2020*

Citation:
Leivada E and Westergaard M (2020) Acceptable Ungrammatical Sentences, Unacceptable Grammatical Sentences, and the Role of the Cognitive Parser. Front. Psychol. 11:364. doi: 10.3389/fpsyg.2020.00364

INTRODUCTION

Introspective linguistic judgments about the well-formedness of linguistic stimuli have long been regarded as one of the most important sources of evidence in linguistics, essentially forming its empirical base (Wexler et al., 1975; Carr, 1990; Schütze, 1996/2016; Baggio et al., 2012). Both the techniques used to elicit such judgments (e.g., controlled experiments, self-introspection, or targeted questioning about whether a specific sentence sounds fine in a specific language) as well as the type of sample that is necessary for the results to have ecological validity (e.g., a pool of participants that is randomly selected from the targeted linguistic community, a non-random sample, or self-introspection) are a matter of debate (see Phillips, 2009; Gibson and Fedorenko, 2010; Sprouse and Almeida, 2013; Branigan and Pickering, 2016). On the other hand, no controversy exists over the fact that judgments about what forms part of a person's linguistic repertoire constitute a rich source of information in theoretical and experimental linguistics.

Since these judgments have such a key role in the study of language, one would expect that the question of *what* they tap into would be one of the first questions in linguistics to provide an indisputable answer to. But that does not seem to be the case. If one searches PubMed or any other database for the terms "acceptability judgment tasks" and "language" on the one hand, and "grammaticality judgment tasks" and "language" on the other, one will quickly discover that the relevant experiments that will show up are the same. They all report findings that are based on the exact same elicitation technique. Perhaps the greatest illustration of how the terms "acceptability" and "grammaticality" are used, often without a clear distinction in place, comes from Schütze's (1996/2016) seminal book on linguistic judgments. While the title of the book talks about "*grammaticality judgments* and linguistic methodology," the very first quote given in the 2016 edition of the book is by Bever (1970), who claims that it is simultaneously the greatest virtue and failing of linguistic theory that *acceptability judgments* are used as the basic data (Schütze, 1996/2016: v). In the preface of the first edition, it is argued that "[t]hroughout much of the history of linguistics, judgments of the *grammaticality/acceptability* of sentences (and other linguistic intuitions) have been the major source of evidence in constructing grammars" (p. xi, emphasis added).

Just as linguists and other cognitive scientists have at times used the terms "ungrammatical" and "unacceptable" roughly synonymously, plurality and overlapping may characterize the use of symbols like ?, *, or ??, that are employed to signal some deviant property of the linguistic stimulus (Bard et al., 1996). To define the relevant terms, the *grammaticality* of a sentence refers to whether the sentence conforms to the syntactic rules of a given language (Fromkin and Rodman, 1998: 106), or put another way, "it is a characteristic of the stimulus itself" (Bard et al., 1996: 33). With respect to *acceptability*, the focus moves from the stimulus to a speaker's perception; in Bard et al.'s (1996) words, it "is a characteristic of the stimulus as perceived by a speaker" (p. 33). Linguistics, however, is not a science that works exclusively with visible primitives; we cannot zoom in on a linguistic stimulus until we find and tease apart an independent, self-contained grammatical core. This means that grammaticality, as one of the possible elements that determine acceptability, "is not directly accessible to observation or measurement" (Lau et al., 2016: 3). The question thus becomes: How do we know anything about grammaticality aside of the information provided by acceptability? Put differently, if grammaticality is defined as "conforming to the rules of the grammar of language X" and if the grammar of language X has the shape that its speakers' judgments and actual performance give it, what way do we have to capture grammaticality other than the one that goes through speakers' perception of well-formedness (i.e., acceptability)?[1]

Answering this question is the main goal of the present work. The starting point of the discussion is Chomsky's (1965) distinction between the terms "acceptability" and "grammaticalness," according to which these two notions and their scales might not coincide, hence his reference to "unacceptable grammatical sentences": sentences that do not form part of grammar for reasons that have nothing to do with grammar. The second aim of the present work is to chart the variation space that is created when one disentangles the two notions: unacceptable grammatical sentences, acceptable ungrammatical sentences, their respective parsability, and the process of assigning them meaning. Last, the scales of grammaticality and acceptability will be discussed and it will be shown that they do not coincide: there are *n* ways of unacceptability, but only two ways of ungrammaticality, in the absolute and the relative sense.

ACCEPTABLE UNGRAMMATICAL SENTENCES AND UNACCEPTABLE GRAMMATICAL SENTENCES

Humans are surprisingly good at providing accurate and consistent judgments about what forms part of their linguistic repertoire.[2] Although informants' opinions about their linguistic behavior are not always concordant with the way they actually speak (Labov, 1996; Cornips and Poletto, 2005), acceptability judgment tasks are reliable as a tool, and the majority of linguistic stimuli can receive unambiguous, consistent judgments. For example, little debate would occur among native speakers of English about the acceptability of (1) or the unacceptability of (2). The former is a grammatically well-formed sentence of English, while the latter is a word-salad that would probably be read and parsed in a rhythm that pertains more to lists of objects than to connected speech.

(1) John said to Mary that he likes doing linguistics.
(2) *To he likes that linguistics John Mary doing said.

Yet, even though such judgments are largely coherent with the actual shape of speakers' internalized grammar, there are some stimuli that have the ability to trick the cognitive parser into unlawfully accepting or rejecting them. Chomsky's (1965) discussion of "unacceptable grammatical sentences" mentions several performance-associated factors that explain why a linguistic stimulus that does not violate any rule of grammar would be rejected by speakers as unacceptable. Factors such as memory limitations, processing constraints, as well as discourse, intonational and stylistic factors may all induce such an effect. For

[1] An obvious answer could be that rules of grammar could be extrapolated through corpora of naturalistic speech. Although such corpora are useful, they cannot substitute judgments, for two reasons. First, they are informative only about what is part of a language, but cannot show the actual limits of variation. It is impossible to establish what is *not* licit in a language only by analyzing them (Henry, 2005). Second, big corpora with rich data that include a variety of genres are the only ones that can provide a faithful approximation of the actual variation space of a language, and these are available only for big, standard languages. This is one of the most important challenges that linguists working with small or non-standard languages face (Leivada et al., 2019a). For these reasons, native judgments are an indispensable tool for most linguists.

[2] This is important because accuracy and stability of judgments are not present in *all* types of judgments that are related to some aspect of human perception. For example, in the famous "The Dress" photograph, not only did judgments of color perception differ across people, with some seeing the dress as blue/black and others as blue/brown or white/gold, but also test-retest reliability revealed switches in perception across testing sessions (Lafer-Sousa et al., 2015).

example, overloading memory and processing resources through nested hierarchies (3) may lead the cognitive parser to not fully register or retain all the relevant information (Gibson and Thomas, 1999), something that is necessary in order to provide an acceptability judgment that faithfully represents whether the stimuli fall inside or outside the domain of predictions of the underlying, internalized grammar. In other words, precisely because of the high complexity of some stimuli, and due to the fact that the cognitive parser works on the basis of processing heuristics (Kahneman, 2011), some deviations may go unnoticed. One such example is (4), which looks very similar to (3) but–unlike (3)–violates a rule of grammar.

(3) The patient the nurse the clinic had hired admitted met Jack. Frazier (1985).
(4) *The doctor the nurse the hospital had hired met John. Frazier (1985).

In linguistic terms, the fact that (4) is missing a verb and has an argument (i.e., "the doctor") that is not assigned any thematic role entails a violation of Chomsky's (1981) θ-criterion, according to which each argument bears one and only one θ-role, and each θ-role is assigned to one and only one argument. Despite the seriousness of this deviation, the "missing verb effect" showed in (4) has been linked to high acceptability rates, even though the sentence is most definitely ill-formed from a syntactic point of view (Gibson and Thomas, 1999). Moreover, this effect is neither restricted to one language nor is it a laboratory phenomenon that arises only in acceptability judgment tasks (Häussler and Bader, 2015). Sentence (4) shows that ease of parsability may influence judgments, and in this specific case, low parsability leads to not spotting a violation of a core syntactic principle. At the same time, high parsability does not guarantee acceptability or grammaticality. For example, speakers of English recognize that (5) expresses a thought that their cognitive parser can easily process, but their language does not produce it in this way.

(5) *What did Peter eat ravioli and?

It seems that a dissociation is in place, because being grammatical (i.e., not violating a rule of grammar) does not guarantee acceptability either. Example (6) is in fact an unacceptable grammatical sentence.[3] Speakers would not judge it as acceptable as (1), but it *is* a grammatically well-formed sentence of English, in the sense that no rule of grammar is violated. Its structure is analogous to that of (7).

(6) Dogs dogs dog dog dogs. Barton et al. (1987).
(7) Cats (that) dogs chase love fish.

The difficulty of (6) suggests that the types of structures that are actually attested in language are influenced by biases of general cognition. One such bias seems to underlie the unacceptability of (6): Identity Avoidance holds that elements of the same phonological and/or syntactic type are unlikely to occur in immediately adjacent positions (van Riemsdijk, 2008). Although this has long been treated as a linguistic ban, recent work has suggested that it has deeper cognitive roots, and more specifically, that it derives from the parser's preference to avoid tokenizing multiple, adjacent occurrences of the same type because of a general bias to provide more attentional resources to novel information ("Novel Information Bias"; Leivada, 2017). Acceptability is thus affected by a variety of processing factors and cognitive biases, and so is grammaticality. For example, although data that flout Identity Avoidance exist [(6); see Leivada, 2017 for examples of syntactic violations], there are no grammatically licit structures that feature five identical, adjacent complementizers, and the prediction is that such structures will never be in use, because a grammar would never consistently deploy them. Even if grammars were able to generate a sentence like *"John said that that that that that Mary kissed him," cognitive biases would intervene and break this sequence of complementizers, for this degree of repetition would not be informative, and by means of looking like noise to the parser, it would make communication infelicitous. A similar situation arises with sentence (4): it is extremely unlikely that a language will consistently deploy sentences with missing verbs that have licensed arguments. In other words, although the rules of the grammar of a language are subject to change in a way that may legitimize the use/acceptability of a previously ill-formed sentence and/or diminish the use/acceptability of a previously attested one, certain changes are not expected to occur, because they violate either a core principle of linguistic cognition or a general cognitive bias.

Talking about a dissociation of acceptability and grammaticality, unacceptable grammatical sentences are one logical possibility. One may wonder whether the other possibility is also attested, i.e., acceptable ungrammatical sentences. Example (8) in **Table 1** provides the missing piece of this dissociation (see also Ross, 2018 for the interaction of grammaticality and acceptability).

Sentence (8) instantiates a linguistic illusion called "comparative illusion" (Montalbetti, 1984). These sentences are called illusions because they trick the parser in a way that renders high acceptability ratings in experiments, even though the stimuli are ill-formed (Wellwood et al., 2018). In linguistic terms, (8) is ill-formed because the main clause subject calls for a comparison of cardinalities of sets, but in the absence of a bare plural in the embedded clause subject, no comparison set is made available (Phillips et al., 2011; O'Connor et al., 2012; Wellwood et al., 2018). Linguistic illusions are the outcome of a partial-match strategy that is operative during processing (Reder and Kusbit, 1991; Kamas et al., 1996; Park and Reder, 2004).

TABLE 1 | A dissociation of grammaticality and acceptability.

	Unacceptable	Acceptable
Grammatical	(6) Dogs dogs dog dog dogs. Barton et al. (1987).	(1) John said to Mary that he likes doing linguistics.
Ungrammatical	(2) *To he likes that linguistics John Mary doing said.	(8) *More people have been to Russia than I have. Montalbetti (1984).

Sentences that were already introduced above appear with their original numbering.

[3]"Dog" can be both a verb and a noun in English. The sentence means the following: dogs that are followed by dogs follow themselves other dogs.

When the parser receives a linguistic stimulus, its components, concepts, and structure are matched to stored knowledge, so that an output is produced. However, the parser matches the stimulus to stored information only up to a point. In other words, a processing threshold is set and the stimulus is checked up to this threshold, hence the notion of partial matching. Given that (8) makes use of locally coherent templates (Townsend and Bever, 2001) that provide a "good-enough fit" (Ferreira and Patson, 2007) for the parser, its ill-formedness may go unnoticed, and this results in high acceptability. Evidently, the way the parser works–via the use of processing heuristics–mediates one's access to the internalized knowledge of grammar. Yet, the ease with which a sentence is unambiguously parsed is not a guarantee for either grammaticality or acceptability. **Table 2** adds high/low parsability to the previous dissociation between grammaticality and acceptability. Once again, all logical possibilities are attested.

Example (9) does not violate any rule of grammar, however, its acceptability is not comparable to that of (1) for semantico-pragmatic reasons that boil down to difficulties that arise "in assigning a coherent meaning to the whole" (Adger, 2018: 161). Unlike (2) or even (10), (9) can be easily parsed in a way that pertains to connected speech. Moreover, a coherent interpretation of it can be provided, and over the years there have been various proposals that construe meanings for it.[4] Perhaps green ideas refer to environmental considerations. One could build a metaphorical narrative where these ideas are colorless and sleeping because at present there is not enough effort to combat climate change, however, their sleep is furious, something that may suggest that some promising initiatives for change are under way. Creating the right context can improve the acceptability of (9) precisely because of its grammatical well-formedness and high parsability.

Perhaps the most interesting sentence of **Table 2** is (4): a sentence that is both ungrammatical and hard to parse, yet still acceptable. Its low parsability hides the grammatical violation, something that leads to high acceptability. Of course, one could claim that such a sentence, despite being labeled "acceptable," would never be attested in one's linguistic performance. However, ungrammatical sentences that are harder to parse *are* in fact attested in naturalistic speech (12a), and the relevant data also include missing verbs in cases of center-embedding (12b).

(12a) "And since I was not informed–as a matter of fact, since I did not know that there were excess funds until we, ourselves, in that checkup after the whole thing blew up, and that was, if you'll remember, that was the incident in which the attorney general came to me and told me that he had seen a memo that indicated that there were no more funds."[5] President Ronald Reagan, April 28, 1987.

(12b) That we scrutinize is a simple consequence of the fact that none of the predictions that you **Δ** during the months that you have been in office has turned out to be true. Häussler and Bader (2015: 14).

Going back to the rest of the data in **Table 2**, we see that (5) and (11) suggest that certain ungrammatical sentences can be easily parsed too. Recent research has suggested that not all ungrammatical sentences receive unclear and unreliable interpretations across speakers (e.g., Etxeberria et al., 2018 on negation). Talking about ungrammatical sentences that are acceptable and parsable, Otero (1972) reached the conclusion that wide acceptability is not a guarantee for grammaticality. Even sentences that have been described as blatantly ungrammatical may actually be acceptable to some degree, and this degree varies across speakers of the same language that have different developmental trajectories (e.g., late bilinguals, heritage speakers, L1 attriters). For example, (5) is ungrammatical because extraction out of coordinated structures is prohibited. A similar island effect has been described for extraction out of relative clauses (13).

(13) *Who do you like the poem that____wrote?

Although much literature portrays such sentences as universally ungrammatical (see Phillips, 2013 and references therein), not all speakers find such violations fully unacceptable. For instance, Lowry et al. (2019) found surprising rates of acceptability for five different types of island violations–including relative clause islands that received a mean score of 3.6 in an 1–5 scale, where 1 stood for the sentence sounding perfectly natural–among late bilingual and heritage speakers of Spanish. Importantly, the two groups differed both in terms of their judgments and in terms of their involuntary physiological reactions that can be proxies for processing effort. In Lowry et al. (2019) these were measured through a pupillometry study: pupil dilation in the ungrammatical stimuli was observed only in the group of late bilinguals, while there was no effect of ungrammaticality in the heritage group. These results suggest that regardless of what a theory/grammar presents as ungrammatical, speakers may successfully parse ungrammatical stimuli in a way analogous to their grammatical counterparts. However, it is an important question whether the parsing is

[4]https://www.physicstomato.com/colorless-green-ideas-sleep-furiously/

[5]http://www.reagan.utexas.edu/archives/speeches/1987/042887e.htm

TABLE 2 | A dissociation of grammaticality, acceptability, and parsability.

	High parsability	Low parsability
Grammatical/ acceptable	(1) John said to Mary that he likes doing linguistics.	(3) The patient the nurse the clinic had hired admitted met Jack. Frazier (1985).
Grammatical/ unacceptable	(9) Colorless green ideas sleep furiously.[1] (Chomsky, 1957)	(10) That that that Bill left Mary amused Sam is interesting is sad. Hornstein (2013).
Ungrammatical/ unacceptable	(5) *What did Peter eat ravioli and?	(2) *To he likes that linguistics John Mary doing said.
Ungrammatical/ acceptable	(11) *Fewer people have been to Tromsø than I have.	(4) *The doctor the nurse the hospital had hired met John. Frazier (1985).

Sentences that were already introduced above appear with their original numbering. [1]*The grammaticality of (9) is fairly indisputable (Hill, 1961), however, not everybody agrees on the degree of its unacceptability. Some scholars have talked about doubtful acceptability, marking the sentence with a "?" to indicate this (Armstrong, 2005), while others have described it as outright unacceptable (Bauer, 2014).*

complete, in the sense that these speakers assign meaning to these ungrammatical stimuli.

Understanding the process of assigning meaning is important in the context of disentangling the role of the parser in acceptable ungrammatical sentences. To illustrate this, let's consider the comparative illusions in **Tables 1**, **2** [examples (8) and (11), respectively], which are ungrammatical but trick the parser into acceptability (Wellwood et al., 2018; Leivada et al., 2019b). Although various experiments have shown that these sentences are assigned a high acceptability rating, one could say that this does not entail that these sentences are actually parsed, in the sense that speakers actually assign them a meaning *m*. A clear exposition of this point is given by Tim Hunter as a reply to Hornstein (2013), who suggests that such sentences may sound good to speakers, but when you ask the people that gave them a high rating what the uttered sentence means, they are unable to provide a meaning:

> *I don't think there is any meaning m such that ("More people have been to Russia than I have," m) is judged acceptable. What* ***is*** *true about these examples is that if you ask whether the string is acceptable without providing any intended interpretation–roughly, if you ask a question of the form "Is there a meaning m such that (s, m) is acceptable?"–then people tend to say "yes." This despite the fact that, as everyone points out, if you ask which meaning this is, people are stumped. [. . .] Why they should make this kind of mistake (i.e., accept the sentence), I have no idea: presumably the answer might be something like, they start searching for a meaning for the string, and they get close enough to feel confident that a meaning* ***can*** *be found without getting all the way there, so they stop and answer "yes" (since no one is asking for the particular meaning).*

In contrast, we suggest that illusions like (8) and (11) are parsed in a way that does go through assigning *m* to *s*. In our work on comparative illusions (Leivada et al., 2019b) we obtained ample evidence that most speakers that judged (8) as acceptable, truly construed an interpretation for it. Among the ones more frequently given by the speakers we tested are: (i) more people than just me have been to Russia, (ii) people have been to Russia more times than I have, and (iii) many people have been to Russia more times than I have (see also Wellwood et al., 2018). Naturally, this is not what the sentence says, but nevertheless, a meaning *is* assigned to the sentence. Also we suggest that one should not ignore the possibility that those speakers that seem stumped upon being asked to provide an interpretation do not do so because they never actually established an association (*s*, *m*), but because in their attempt to articulate the latter, they spot the illusion. Crucially, this does not entail that at no point were they actually able to put their finger on a possible meaning.

The second interesting issue with Hunter's point has to do with the juxtaposition of two very different ways of eliciting judgments through asking "Is *s* acceptable?" or "Is there a meaning *m* such that (*s*, *m*) is acceptable?" These two questions do not tap into the same thing. Previous research on the pragmatics of cognitive illusions has proposed that when processing such sentences, the hearer searches for meaning within a manipulative communication, that is, within a tricky context that features a "manipulation (that) can be best defined in terms of the constraints it imposes on mental processing" (Maillat and Oswald, 2009: 361). In this context, the hearer stops searching for meaning after finding one that sufficiently meets her expectations of relevance in accordance with the previous discourse. The illusion thus arises in the process of selecting meaning within a manipulative context that takes advantage of (i) the parser's limitations and (ii) the parser's way of operating through employing certain processing heuristics such as partial matching or shallow processing.

If relevance and previous context can bias an acceptability judgment through creating the necessary conditions for an illusion to arise, the bias will be even greater if a specific *m* is given to a participant point-blank in an experiment that asks "Is there a meaning *m* such that (*s*, *m*) is acceptable?" As shown in Tversky and Kahneman's (1974) work, options in a task are evaluated relative to some *reference point*. Theoretically speaking, the reference point in standard acceptability judgment tasks is the linguistic repertoire of the tested speaker: We often instruct speakers to disregard the formal prescriptive rules of grammar and focus on evaluating the stimuli on the basis of how they use the language. If we add a given *m* to this picture, we alter the reference point. This does not mean that such a task cannot provide useful and informative findings, but that possibly the obtained findings will not be tapping directly into a speaker's perception of her idiolect. Instead, it will be mediated by an *anchoring effect* that may cause an *adjustment* to the speaker's judgment on the basis of *m*. To understand this effect, consider the following example by Kahneman (2011).

(14a) Was Gandhi more or less than 144 years old when he died?
(14b) How old was Gandhi when he died? Kahneman (2011: 122).

Of course nobody claimed that Gandhi was 144 years old when he died, but it has been found that when (14b) is presented after (14a), the provided high number functions as an anchor that affects people's estimate (Kahneman, 2011). To draw the analogy with judgment tasks, let's compare (15a) to (15b), and it will become clear why "Is *m* acceptable?" does not ask the same thing as "Is there a meaning *m* such that (*s*, *m*) is acceptable?."

(15a) Assuming a scale from 1 to 5, how acceptable is *s* on the basis of an intended meaning *m*?
(15b) Assuming a scale from 1 to 5, how do you rate *s* on the basis of your idiolect?

In (15a), the possibility of *s* getting a meaning is explicit and a possible meaning *m* is already given to the speaker as part of the question that introduces the stimuli *s*. This can bias the rating of *s* on the basis of the "anchor-and-adjust" heuristic.

To sum up, illusions do not necessarily entail that parsing fails to produce a meaning, but that the parser can be tricked into providing both a meaning and an acceptability rating that may not correspond to the actual status of the stimulus in terms of what the speaker's internalized grammar looks like. Importantly, a number of factors contribute to this process of tricking the parser: context, task and stimuli presentation, as well as structural complexity are only a few.

The relation between grammatical well-formedness and acceptability is a complex one. As mentioned in the Introduction, the main goal of the present work is to discuss whether acceptability is an indispensable gateway to grammaticality or whether there is a way of capturing grammaticality other than the one that goes through speakers' perception of what is well-formed in their native linguistic repertoire (i.e., acceptability). Having presented the dissociation between acceptability, grammaticality and the way the parser works, the next section deals with how grammaticality is established and where it comes from.

WHERE DOES GRAMMATICALITY COME FROM?

Asking about the origin of the rules of grammar, Adger (2019) suggests that we learn them: They come from the way people speak. Although this is true, the issue is more complex, because different people speak in different ways even within linguistic communities that feature only one language. When one says that (5) and (8) are ungrammatical in English, this use of the term "ungrammatical" is not meant to be interpreted as a faithful representation of every English speaker's idiolect in an individual way, precisely because even monolingual speakers in a monolingual community show variation.[6] Rather it refers to some established consensus about what is the norm in a specific variety of English; a norm that the grammar books describe in detail. Put differently, if some speakers of English, Spanish, or German accept to some degree or even produce to some degree island violations (Lowry et al., 2019 for Spanish), missing verbs in nested hierarchies (Häussler and Bader, 2015 for English and German), or comparative illusions[7], do we want to say that these structures are grammatical in English, Spanish, and German? While it certainly appears to be the case that some speakers' grammars may occasionally give rise to such structures, we should take into account that, in relation to naturalistic data, production factors may endow the linguistic message with noise (i.e., false starts, infelicitous lemma retrievals, missing elements due to memory constraints, etc.), which can account for how some of these ungrammatical sentences come to be produced in spontaneous speech. In relation to the possible acceptability of these structures in experimental settings, the previous section has shown that there is a dissociation between acceptability and grammaticality, such that we should expect some degree of discrepancy between the way speakers judge sentences in an experiment (where even the way the stimuli are presented may influence judgments; see examples 14–15), the way they actually speak, and the way that prescriptive grammar says they (should) speak.

The question still holds: Where does grammaticality come from? The tentative answer we offer is that grammaticality is often a formal, standardized snapshot of the way the official language looks like at a given point in time. Grammaticality is constantly redefined through ever-changing acceptability, but it also reflects stable properties of general cognition. In this context, we do not know much about grammaticality outside acceptability (recall that observation of naturalistic data cannot reveal what is ungrammatical in a language) in the sense that there is no list of grammatical properties that are grammatical *in and of themselves*. They are all grammatical within a context that is called language X. Language X is constantly changing and what is (un)grammatical today may not be (un)grammatical tomorrow, depending on whether the new speakers of X find it acceptable or not and whether this acceptability is generalized and established as the norm or not. For example, Ancient Greek featured a syntactic phenomenon called Attic syntax which permitted a number mismatch between the plural, neuter subject and the verb (16a). This structure is not a grammatically licit option in Modern Greek (16b), but not because there is something intrinsically *un*grammatical about it; it simply does not form part of the grammar anymore. Phrased differently, there is no notion of *self-contained* grammaticality that (16a) has and (16b) lacks; they just form part of two different snapshots of a grammar's domain of predictions at different points in time.

(16a) Ta padia pezi. [Ancient Greek]
the child.PL play.3SG
"The children are playing."

(16b) *Ta koritsia gela. [Modern Greek]
the girl.PL laugh.3SG
Intended meaning: "The girls are laughing."

This claim is partially concordant with Chomsky et al.'s (2019); see also Chomsky (1993) view that in natural language there exists no *independently given* notion of grammatical well-formedness. Indeed, the grammatical well-formedness of a linguistic stimulus does not boil down to an independently definable grammatical core, but is a mere historical "accident" that (i) refers to whether the stimulus forms part of the standardized snapshot or not and (ii) is subject to change such as the one shown in (16a-b). Nevertheless, this view is true only for one reading of the term "grammatical": grammatical as *actually forming part of the grammar of a specific language*.

However, we suggest there is also another reading of the term "grammatical." To understand this other reading, one needs to factor in that change is not without limits. Not all changes are possible and not all linguistic stimuli are candidates for forming part of grammar. For example, as mentioned in the

[6]For example, Smith and Cormack (2002) discuss sequences of tense possibilities in English. With some speakers accepting "Did you know that Emily is ill?" and with others considering it unacceptable (i.e., accepting only "Did you know that Emily was ill?"), these authors capture the observed variation by suggesting that this is "a situation in which intuitions are completely clear-cut, so the relevant parameter has been fixed, but it has been fixed apparently at random, presumably because of the paucity of distinguishing data" (p. 286). Another example is given in Levelt (1972), who showed that opinions about what is grammatical in a language are not uniform even among trained linguists who are native speakers of the language in question. When he asked 24 linguists to judge whether the sentence "The talking about the problem saved her" (Fraser, 1970, p. 91, with the example marked as ungrammatical) was marked as grammatical or ungrammatical in a specific linguistics article, he found that judgments varied, and only 1/3 of the consulted linguists gave the judgment "ungrammatical," in agreement with the original source.

[7]One example of a comparative illusion in naturalistic speech, outside of an experimental setting, is the following tweet by Dan Rather: "I think there are more candidates on stage who speak Spanish more fluently than our president speaks English." [Available at https://twitter.com/danrather/status/1144076809182408704]

section "Acceptable ungrammatical sentences and unacceptable grammatical sentences" there are no grammatically licit structures that feature five identical, adjacent complementizers, and the prediction is that such structures will never be grammatical. Similarly, a sentence such as (4), which violates the θ-criterion, is unlikely to ever form part of grammar.[8] As discussed in the next section, certain changes are not expected to occur, because they violate either a core principle of language (e.g., the θ-criterion in 4) or a general cognitive bias (e.g., the Novel Information Bias in 6). In this sense, Chomsky et al. (2019) are right in arguing that there exists no independently given notion of grammatical well-formedness, but we would like to add to their claim that there do exist independently given constraints to the set of entities that this notion can encompass. This is the other reading of the term: grammatical *as having the potential to be a part of grammar*, by means of not going against any of the relevant biases and communication/processing principles that underlie language and cognition.

N TYPES OF UNACCEPTABILITY AND TWO TYPES OF UNGRAMMATICALITY

It is an uncontroversial claim that acceptability judgments are not categorical, but form a continuous spectrum (Sprouse, 2007 and references therein). The usual meaning of the word "continuous" is *unbroken* or *undivided*, hence it is the nature of a continuum to be undivided, or better, to permit repeated division without limit (Bell, 2017). If one subscribes to the view that acceptability should be viewed as a continuum, one also subscribes to the view that the acceptability continuum is *infinitely divisible*. Although acceptability judgment tasks that involve Likert scales feature a finite number of options more often than not, there are experiments that ask speakers to judge a linguistic stimulus by adjusting a slider on a continuum without any clearly delineated categories such as "acceptable," "somewhat acceptable" etc.

While the scale of unacceptability involves *n* positions, the scale of ungrammaticality involves only two: Something can be ungrammatical in the *relative* or in the *absolute* sense. The relative sense pertains to the first reading of the term "grammatical" that was mentioned above: forming part of grammar. We call it "relative" because it is defined in the context of a given language. For example, (16b) is ungrammatical *in relation to* Modern Greek, but it is not ungrammatical *per se*. It is a potential candidate for forming part of grammar, it was grammatical in the past (16a), and it may be again in the future. Similarly, (17) is ungrammatical in relation to Standard English, but this is an accident, as it could potentially be grammatical (and in fact it is grammatical in many varieties of English, including e.g., Belfast English; Henry, 2005).

(17) The children is here.

Relative ungrammaticality (i) is subject to change, (ii) is defined in the context of a specific language, and (iii) refers to those sentences that could be grammatical, but for some reason are not in the language in question, yet they probably are in some other language. Absolute ungrammaticality (i) is not subject to change, (ii) is not defined in relation to one given language, and (iii) concerns violations of some core principle of language and/or cognition, that is, structures that grammar would never consistently deploy. Therefore, absolute ungrammaticality has to do with structures that *cannot* form part of grammar.

Comparing the scales of the two notions, acceptability and grammaticality, it is meaningful to talk about partial acceptability (Sprouse, 2007), but not about partial or strong-weak ungrammaticality. A rule of grammar (or more than one rule of grammar) can be either violated or not, but it cannot be violated just a bit. Ungrammaticality cannot be a matter of degree, only acceptability can. Put differently, a native speaker can judge a structure in her language as more acceptable than another structure, but a structure forming part of a grammar cannot be more grammatical than another structure that forms part of the same grammar.

Although some scholars have talked about "partial ungrammaticality," we would argue that this refers either to partial unacceptability or to variation in a linguistic community. Consider, for instance, the discussion of partial ungrammaticality in Attinasi (1974): "A hidden assumption of homogeneity, that the language competence of every speaker consists of the same structures, falters when the question of partial ungrammaticality is raised. How can some speakers totally reject, others partially accept and still others totally accept certain sentences as grammatical if each presumably speaks 'English,' or any other language?" (p. 280). In our view, this question has to do with gradient *acceptability*: that is what speakers have judgments about.[9] As we have seen, grammaticality can be dissociated from acceptability. Also, the observed variation does not entail or legitimize the notion of partial grammaticality, because, as mentioned in the section "Where does grammaticality come from?," different people speak in different ways, but grammaticality evokes an established norm that is part of a formal snapshot. Speakers may deviate from this norm, either because language change has occurred and the norm does not reflect this yet, or because their idiolect simply differs from the norm. But this should be referred to as "interspeaker variation," not "partial grammaticality."

OUTLOOK

The present work has discussed the complex relation between grammaticality, acceptability, and parsability. A number of unacceptable grammatical sentences and acceptable ungrammatical sentences have been presented, including grammatical illusions, violations of Identity Avoidance, and sentences that involve a high level of processing complexity

[8] Although the missing-verb effect can be occasionally attested in naturalistic speech (12b), we argue that this has to do with production factors that introduce noise to the linguistic message.

[9] Boeckx (2010) rightly calls the term "grammaticality judgment tasks" a misnomer, because speakers lack intuitions about whether something is grammatical. In the absolute meaning of the term "grammatical," having judgments about grammaticality would entail having intuitions about the workings of *all* linguistic and cognitive factors that determine the limits of grammar, and no speaker (or linguist for that matter) has that.

that overloads the cognitive parser. Focusing on acceptable ungrammatical sentences, we have argued that in many cases their acceptability entails that a meaning has been assigned to them. Also, two notions of ungrammaticality have been introduced: (Un)grammaticality in the relative sense refers to the whether the stimulus falls within the domain of predictions of a given grammar or not. (Un)grammaticality in the absolute sense refers to whether the stimulus has the potential to be a part of grammar or not. Relative (un)grammaticality is an ever-changing property of the stimulus, whereas absolute (un)grammaticality is stable. In both readings of the term, grammaticality is defined by something that is external to the stimulus (be it the grammar of a specific language or principles of general/linguistic cognition), and it is not an inextricable property of the stimulus itself. Put differently, there is no list of properties that are (relatively/absolutely) grammatical *in and of themselves,* or as Chomsky et al. (2019) phrase it, there is no independently given notion of grammatical well-formedness in natural language.

Through disentangling the various uses of the terms "acceptable" and "grammatical," the overarching aim of this work has been to aid the field in reaching a more adequate level of terminological clarity for notions that pertain to the evidential base of linguistics. Many details of the distinction between relative and absolute (un)grammaticality are left to be worked out, and this will likely be the topic of future work. To give just one example, when we deal with island effects of the sort discussed above, do we deal with absolute ungrammaticality that is universal and derives from processing or other principles of language/cognition, or with relative ungrammaticality that is manifested in different ways across different languages, precisely because it is defined on the basis of language-specific factors? Or as Ott (2014): 290) asks is "*What does John like and oranges?" ungrammatical (in the absolute sense that *it cannot be generated by the grammar*), given that speakers can easily assign it a transparent interpretation (e.g., which *x*: John likes *x* and oranges)? The answer is currently unclear to us, and it probably needs novel experimental work to be properly discussed. Recognizing this uncertainty does not mean undermining the proposed distinction between absolute and relative ungrammaticality. It rather suggests that progress is underway, or as (Feynman, 1998: 27) puts it, "[b]ecause we have the doubt, we then propose looking in new directions for new ideas. The rate of the development of science is not the rate at which you make observations alone but, much more important, the rate at which you create new things to test."

AUTHOR CONTRIBUTIONS

EL and MW conducted the research behind this work. EL drafted a first manuscript, which MW revised. Both authors contributed equally to the final editing of this work.

FUNDING

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant Agreement No. 746652. The publication charges for this manuscript have been funded by a grant from the publication fund of UiT The Arctic University of Norway. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Adger, D. (2018). "The autonomy of syntax," in *Syntactic Structures After 60 Years: The Impact of the Chomskyan Revolution in Linguistics*, eds N. Hornstein, H. Lasnik, P. Patel-Grosz, and C. Yang, (Berlin: Mouton de Gruyter), 153–176.

Adger, D. (2019). Where do the rules of grammar come from? *Psychol. Today*. Available online at: https://www.psychologytoday.com/us/blog/language-unlimited/201908/where-do-the-rules-grammar-come (accessed February 27, 2020).

Armstrong, N. (2005). *Translation, Linguistics, Culture*. Clevedon: Multilingual Matters.

Attinasi, J. T. (1974). The sociolinguistics of William Labov. *Biling. Rev.* 1, 279–304.

Baggio, G., van Lambalgen, M., and Hagoort, P. (2012). "Language, linguistics and cognition," in *Handbook of the Philosophy of Science: Philosophy of Linguistics*, Vol. 14, eds R. Kempson, T. Fernando, and N. Asher, (Amsterdam: Elsevier), 325–355. doi: 10.1016/b978-0-444-51747-0.50010-x

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language* 72, 32–68.

Barton, G. E., Berwick, R. C., and Ristad, E. S. (1987). *Computational Complexity and Natural Language*. Cambridge, MA: MIT Press.

Bauer, L. (2014). Grammaticality, acceptability, possible words and large corpora. *Morphology* 24, 83–103. doi: 10.1007/s11525-014-9234-z

Bell, J. L. (2017). "Continuity and Infinitesimals," in *The Stanford Encyclopedia of Philosophy*, *Summer 2017 Edition*, ed. E. N. Zalta. Available online at: https://plato.stanford.edu/archives/sum2017/entries/continuity/ (accessed February 27, 2020).

Bever, T. G. (1970). "The influence of speech performance on linguistic structure," in *Advances in Psycholinguistics*, eds G. B. Flores d'Arcais, and W. J. M. Levelt, (Amsterdam: North-Holland Publishing Co.), 4–30.

Boeckx, C. (2010). *Language in Cognition. Uncovering Mental Structures and the Rules Behind Them*. Maden: Wiley-Blackwell.

Branigan, H. P., and Pickering, M. J. (2016). An experimental approach to linguistic representation. *Behav. Brain Sci.* 40:e282. doi: 10.1017/s0140525x16002028

Carr, P. (1990). *Linguistic Realities: An Autonomist Metatheory for the Generative Enterprise*. Cambridge: Cambridge University Press.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

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

Chomsky, N. (1993). "Aminimalist program for linguistic theory," in *The View from BUILDING 20: Essays in Linguistics in Honor of Sylvain Bromberger*, eds K. Hale, and S. J. Keyser, (Cambridge, MA: MIT Press), 1–52.

Chomsky, N., Gallego, A. J., and Ott, D. (2019). Generative grammar and the faculty of language: insights, questions, and challenges. *Catalan J. Linguist.* Available online at: https://ling.auf.net/lingbuzz/003507 (accessed February 27, 2020).

Cornips, L., and Poletto, C. (2005). On standardising syntactic elicitation techniques (part 1). *Lingua* 115, 939–957. doi: 10.1016/j.lingua.2003.11.004

Etxeberria, U., Tubau, S., Deprez, V., Borràs-Comes, J., and Espinal, M. T. (2018). Relating (Un)acceptability to Interpretation. Experimental investigations on negation. *Front. Psychol.* 8:2370. doi: 10.3389/fpsyg.2017.02370

Ferreira, F., and Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Lang. Linguist. Compass* 1, 71–83. doi: 10.1111/j.1749-818x.2007.00007.x

Feynman, R. P. (1998). *The Meaning of It All. Thoughts of a Citizen Scientist*. Reading: Perseus Books Group.

Fraser, B. C. (1970). "Some remarks on the action nominalization in English," in *Readings in English Transformational Grammar*, eds R. A. Jacobs, and P. S. Rosenbaum, (Waltham, MA: Ginn), 83–98.

Frazier, L. (1985). "Syntactic complexity," in *Natural Language Parsing. Psychological, Computational and Theoretical Perspectives*, eds D. R. Dowty, L. Karttunen, and A. Zwicky, (Cambridge: Cambridge University Press), 129–189.

Fromkin, V., and Rodman, R. (1998). *An Introduction to Language*, 6th Edn. Orlando, FL: Holt, Rinehart and Winston.

Gibson, E., and Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends Cogn. Sci.* 14, 233–234. doi: 10.1016/j.tics.2010.03.005

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cognit. Process.* 14, 225–248. doi: 10.1080/016909699386293

Häussler, J., and Bader, M. (2015). An interference account of the missing-VP effect. *Front. Psychol.* 6:766. doi: 10.3389/fpsyg.2015.00766

Henry, A. (2005). Non-standard dialects and linguistic data. *Lingua* 115, 1599–1617. doi: 10.1016/j.lingua.2004.07.006

Hill, A. A. (1961). Grammaticality. *Word* 17, 1–10. doi: 10.1080/00437956.1961.11659742

Hornstein, N. (2013). *Acceptability and Grammaticality*. Available online at: http://facultyoflanguage.blogspot.com\penalty-\@M/2013/02/acceptability-and-grammaticality.html (accessed February 22, 2013).

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kamas, E. N., Reder, L. M., and Ayers, M. S. (1996). Partial matching in the Moses illusion: response bias not sensitivity. *Mem. Cognit.* 24, 687–699. doi: 10.3758/bf03201094

Labov, W. (1996). "When intuitions fail," in *Papers from the Parasession on Theory and Data in Linguistics Chicago Linguistic Society*, Vol. 32, eds L. McNair, K. Singer, L. Dolbrin, and M. Aucon, (Chicago, IL: Chicago Linguistic Society), 77–106.

Lafer-Sousa, R., Hermann, K. L., and Conway, B. R. (2015). Striking individual differences in color perception uncovered by "the dress" photograph. *Curr. Biol.* 25, R545–R546. doi: 10.1016/j.cub.2015.04.053

Lau, J. H., Clark, A., and Lappin, S. (2016). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cogn. Sci.* 41, 1202–1241. doi: 10.1111/cogs.12414

Leivada, E. (2017). What's in (a) label? Neural origins and behavioral manifestations of identity avoidance in language and cognition. *Biolinguistics* 11, 221–250.

Leivada, E., D'Alessandro, R., and Grohmann, K. K. (2019a). Eliciting big data from small, young, or non-standard languages: 10 experimental challenges. *Front. Psychol* 10:313. doi: 10.3389/fpsyg.2019.00313

Leivada, E., Mitrofanova, N., and Westergaard, M. (2019b). "The impact of bilingualism in processing cognitive illusions," in *Talk at the Capturing and Quantifying Individual Differences in Bilingualism workshop*, (Tromsø: UiT-The Arctic University of Norway).

Levelt, W. J. M. (1972). Some psychological aspects of linguistic data. *Linguist. Ber.* 17, 18–30.

Lowry, C., Madsen, C. N. II, Phillips, I., Martohardjono, G., and Schwartz, R. G. (2019). "Gradience in Spanish island violations: a psychophysiological study of two bilingual groups," in *Proceedings of the Experimental Psycholinguistics Conference*, Palma.

Maillat, D., and Oswald, S. (2009). Defining manipulative discourse: the pragmatics of cognitive illusions. *Int. Rev. Pragmat.* 1, 348–370. doi: 10.1163/187730909x12535267111651

Montalbetti, M. (1984). *After Binding:On the Interpretation of Pronouns*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

O'Connor, E., Pancheva, R., and Kaiser, E. (2012). "Evidence for online repair of Escher sentences," in *Proceedings of Sinn und Bedeutung*, Vol. 17, eds E. Chemla, V. Homer, and G. Winterstein, (Paris: École Normale Supérieure), 363–380.

Otero, C. (1972). Acceptable ungrammatical sentences in Spanish. *Linguist. Inq.* 3, 233–242.

Ott, D. (2014). Syntactic islands by Cedric Boeckx (review). *Language* 90, 287–291. doi: 10.1353/lan.2014.0008

Park, H., and Reder, L. M. (2004). "Moses illusion: implication for human cognition," in *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgment, and Memory*, ed. R. F. Pohl, (Hove: Psychology Press), 275–292.

Phillips, C. (2009). "Should we impeach armchair linguists?," in *Japanese/Korean Linguistics*, Vol. 17, eds S. Iwasaki, H. Hoji, P. M. Clancy, and S.-O. Sohn, (Stanford, CA: CSLI Publications), 49–64.

Phillips, C. (2013). "On the nature of island constraints I: language processing and reductionist accounts," in *Experimental Syntax and Island Effects*, 64-108, eds J. Sprouse, and N. Hornstein, (Cambridge: Cambridge University Press).

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces*, Vol. 37, ed. J. Runner (Bingley: Emerald Publications), 147–180.

Reder, L. M., and Kusbit, G. W. (1991). Locus of the Moses illusion: imperfect encoding, retrieval, or match? *J. Mem. Lang.* 30, 385–406. doi: 10.1016/0749-596x(91)90013-a

Ross, D. (2018). Conventionalization of grammatical anomalies through linearization. *Stud. Linguist. Sci.* 42, 1–28.

Schütze, C. T. (1996/2016). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Berlin: Language Science Press.

Smith, N., and Cormack, A. (2002). Parametric poverty. *Glot Int.* 6, 285–287.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1, 123–134.

Sprouse, J., and Almeida, D. (2013). The empirical status of data in syntax: a reply to Gibson and Fedorenko. *Lang. Cogn. Process.* 28, 222–228. doi: 10.1080/01690965.2012.703782

Townsend, D., and Bever, T. G. (2001). *Sentence Comprehension: Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.

van Riemsdijk, H. (2008). "Identity avoidance: OCP effects in Swiss relatives," in *Foundational Issues in Linguistic Theory: Essays in Honor of Jean-Roger Vergnaud*, eds R. Freidin, C. P. Otero, and M. L. Zubizarreta, (Cambridge, MA: MIT Press), 227–250.

Wellwood, A., Pancheva, R., Hacquard, V., and Phillips, C. (2018). The anatomy of a comparative illusion. *J. of Semant.* 35, 543–583. doi: 10.1093/jos/ffy014

Wexler, K., Culicover, P., and Hamburger, H. (1975). Learning-theoretic foundations of linguistic universals. *Theor. Linguist.* 2, 215–224. doi: 10.1515/thli.1975.2.1-3.215

Conflict of Interest: 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
published: 24 March 2020
doi: 10.3389/fpsyg.2020.00395

Wh-Movement, Islands, and Resumption in L1 and L2 Spanish: Is (Un)Grammaticality the Relevant Property?

*Sílvia Perpiñán**

Department of Applied Linguistics, Universitat Internacional de Catalunya, Barcelona, Spain

This study reflects on the meaning of the results of a self-paced grammaticality judgment task that tested island configurations (with gaps and resumptive pronouns) in L1 and L2 speakers of Spanish. Results indicated that resumptive pronouns do not rescue extractions from islands, as traditionally assumed in grammatical theory, and propose that islands are essentially an interpretative or processing matter, and not only a grammatical one, as in Kluender (1998). This study further challenges the L2 studies that proposed that L2 learners are fundamentally different from native speakers because they usually fail to reject island configurations, and shows that L2 learners are sensitive to the same processing and interpretative mechanisms that native speakers employ to parse island configurations. Generally speaking, this study proposes that apparent purely syntactic restrictions such as extractions from islands might not depend on their grammatical formation, but on other relevant factors such as plausibility, embedding, and processability, which together with grammatical well-formedness configure a more holistic and useful notion of linguistic acceptability.

Keywords: wh-movement, islands, spanish, processibility, L2 learners, resumptive pronoun

OPEN ACCESS

Edited by:
Susagna Tubau, Autonomous University of Barcelona, Spain

Reviewed by:
Eloi Puig-Mayenco, University of Southampton, United Kingdom
Mike Putnam, Pennsylvania State University (PSU), United States

***Correspondence:**
Sílvia Perpiñán
silvia.perpinan@gmail.com

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

Received: *30 September 2019*
Accepted: *20 February 2020*
Published: *24 March 2020*

Citation:
Perpiñán S (2020) Wh-Movement, Islands, and Resumption in L1 and L2 Spanish: Is (Un)Grammaticality the Relevant Property? Front. Psychol. 11:395. doi: 10.3389/fpsyg.2020.00395

INTRODUCTION

The concept of grammaticality has been of vital importance in the development of the field of modern linguistics, particularly since Chomsky's influential books, *Syntactic Structures* (Chomsky, 1957) and *Aspects of the Theory of Syntax* (Chomsky, 1965). The study of what is possible and, crucially, what is not possible in a language has allowed us to deepen our knowledge on particular and universal properties of linguistic systems. In the field of Second Language Acquisition from a Generative Perspective (GenSLA), the notion of grammaticality has also been essential in order to determine the nature of interlanguage grammars and to describe the implicit linguistic knowledge of a second language learner. Generative linguistics generally assumes that Universal Grammar (UG), which is domain-specific, takes care of the breach left between what is acquired through input and what is deduced by general cognition. Much of the debate in GenSLA during the 80s and 90s revolved around whether interlanguage grammars and native grammars are fundamentally similar or fundamentally different, and whether the former could access UG after the critical period of acquisition (for a summary, see White, 1989, 2003). Constraints on *wh*-movement, i.e.: Subjancency, have been taken as the ideal case to test the accessibility of interlanguage grammars to UG since they typically

illustrate the poverty-of-the-stimulus problem. Islands are not present in the L2 input or taught in a classroom setting, and one can find L1 languages in which *wh*-movement does not operate. Therefore, these L2 learners cannot rely on L1 knowledge or direct L2 input to know the restrictions on *wh*-movement. The logic goes as follows: if we can show that these L2 learners whose L1 does not have *wh*-movement obey the subjacency constraints that regulate *wh*-movement, then we can conclude that their knowledge must come from UG (but see Pearl and Sprouse, 2013 for a different explanation). With this in mind, researchers have traditionally employed Grammaticality Judgment Tasks (GJT) as a technique to tap into the underlying grammatical representation of (non-)native speakers, which crucially affords us to test both possible and impossible sentences. This study reflects on the concept of grammaticality in both native and non-native grammars, and on how it has been used to argue for or against the accessibility to UG by adult second language learners, a central issue in GenSLA. It further questions the assumed (un)grammaticality of certain complex structures, such as island configurations or islands rescued by resumptive pronouns (RPs), and particularly, its syntactic nature. Heestand et al. (2011) already proposed that resumption does not necessarily rescue islands in English, but the application of these recent ideas in the second acquisition research has been very scarce, and L2 data that support these claims are practically inexistent. Likewise, a similar study on L2 Spanish is missing. Moreover, the acquisition of oblique relative clauses is widely unexplored, particularly in real-time use, in which processing resources might be compromised and resumption as a last resort could be favored (McCloskey, 1990). The present study aims to fill these gaps in the literature.

THE LINGUISTIC PHENOMENON: ISLANDS AND *WH*-MOVEMENT

Wh-movement is an extensively studied topic in generative linguistics, especially since Chomsky (1977) proposed that the transformation involved in questions, relative clauses, comparatives, or easy-to-please constructions could be reduced to the general "wh-movement" transformation, a successive cyclic movement to COMP. Later, Chomsky (1981)'s Government and Binding framework presented wh-movement as an instance of a more general transformation: *move* α, regulated among others, by the Subjacency Principle (Chomsky, 1986), which basically controls how far a *wh*-phrase can move, and is supposed to be universal. The original subjacency condition posited that "a constituent may not move over more than one bounding category at a time" (Chomsky, 1973). Even though the concept of bounding nodes may have changed as linguistic theory has evolved, the idea is that Subjacency explains the contrast between (1b) and (2b) because in (1b), the *wh*-word crosses one bounding node at a time, first the IP and then the CP, with successive cyclic movement; whereas in (2b), the first movement crosses one bounding node, -the IP-, but it crosses two in the second movement, the CP and the DP, which renders the sentence ungrammatical. This observation led to propose that complex DPs, in this case a Relative Clause, are "islands" [in Ross' (1967) terminology] from which a *wh*-word cannot be extracted. Examples from Belikova and White (2009):

(1) a. You said this girl danced with **John**.
b. **Who**$_i$ did $_{IP}$[you say $_{CP}$[t_i that $_{IP}$ [this girl danced with t_i]]]?

(2) a. You met a girl that danced with **John**.
b. ***Who**$_i$ did $_{IP}$[you meet $_{DP\text{-}RC}$[a girl $_{CP}$[t_i that $_{IP}$ [danced with t_i]]]]?

In the last 20 years, there has been a significant amount of experimental work that aims to explain the source of the unacceptability of *island effects* (see Sprouse and Hornstein, 2013 for a summary), a classic issue in syntactic theory since Ross (1967). Much has been debated regarding whether islands are a grammatical entity or a parsing one; that is, whether the structure-building constraints that restrict *wh*-movement from certain domains are a syntactic grammatical representation in the cognitive system (a position usually termed as "grammatical theories", Phillips, 2013) or whether islands effects arise as a result of a processing failure or processing limitation, an epiphenomena that comprehends multiple factors such as semantic anomaly, processing difficulty, etc. ("resource-limitation theories," Kluender, 1991, 1998; Kluender and Kutas, 1993; Hofmeister et al., 2013; Kluender and Gieselman, 2013). This dichotomy closely ties grammatical theories with real-time language processing (Phillips, 2006; Lewis and Phillips, 2015) and echoes a fundamental controversy in SLA theories when trying to explain the cause of non-convergence in L2 learners (representational vs. computational accounts, Hopp, 2007, 2009; Slabakova, 2009; Perpiñán, 2015). That is, whether L2 learners have permanent representational deficits, probably due to a partial (Hawkins and Chan, 1997), or no access to UG (Bley-Vroman, 1990, 2009; Meisel, 1997), or whether L2 learners are not able to process the language as efficiently or with the same syntactic detail as native speakers (Clahsen and Felser, 2006). Sprouse et al. (2012) even consider, although do not defend, a third option to explain island effects in L1, which is a combination of the grammatical and reductionist accounts, termed *grounded theories*. Grounded theories assume that island effects are caused by grammatical constraints that have been grammaticized over time because if these structures were generated, these would be difficult to parse. To summarize, the island debate in native languages is an especially multifactorial puzzle that adds to the unresolved challenges in the study of L2 knowledge and its processing, the current debates in the field of SLA.

Ever since Subjacency was put forward as a grammatical explanation of island violations, it has been studied widely in the SLA field as it allows us to make pertinent predictions regarding the role of UG in the interlanguage grammar. If L2 acquisition is constrained in all its instances by UG, then, all possible L2 interlanguage grammars should obey universal principles, including the Subjacency Principle, regardless of the learners' L1, the target language, and their *wh*-movement properties. Island configurations have been typically used as a test for syntactic

movement: if an extraction requires syntactic movement, then, that construction will be ungrammatical if it is extracted from an island. If, on the contrary, a constituent is apparently extracted from an island and the derivation is not ruled out, then it is assumed that there was no movement involved. In that case, we would say that there was not an extraction *per se*, but that the constituent was base-generated and bound somehow with its antecedent.

Traditionally, an assumed way to rescue an island violation is by introducing a resumptive pronoun (Ross, 1967; Kroch, 1981; McCloskey, 1990; Shlonsky, 1992). According to McCloskey (2007), we can group three types of languages that employ resumptive pronouns (RPs) differently; in this study we are concerned with two of these types, Type I and Type III. Type I languages would allow free variation of RPs and gaps; inside an island though, only resumptive pronouns can appear. This is the case of Lebanese Arabic as described by Aoun et al. (2001), and we will assume that it is also the case of Moroccan Arabic, the variety that concerns us in this experiment. However, Shlonsky (1992) argues that the use of (true) resumptive pronouns in Hebrew and Palestinian respond to a last resort strategy, meaning that they are used when operations general to Universal Grammar are blocked. According to this author, the use of resumptive pronouns is a language-specific rule that must apply whenever movement is not available, and it is not optional. This might be true for direct object relative clauses, but Arabic prepositional relative clauses present both strategies, movement and resumptive pronouns, as explained below. Type III languages are those that present "intrusive pronouns" (Sells, 1984), which are not a true pronoun or syntactically active resumptive (Asudeh, 2012) as it does not alternate with gaps and is not island-sensitive. We are assuming that this is the case for both English and Spanish.

Recently, there have been different proposals to explain RPs, and their power (or lack thereof) to ameliorate illicit island extractions has been seriously questioned. In a nutshell, syntactic and off-line data seem to indicate that RPs do improve island violations, whereas psycholinguistic data have failed to find strong evidence that supports this claim. For instance, Alexopoulou and Keller (2007), as well as Heestand et al. (2011), and Polinsky et al. (2013), in a series of experimental studies testing different types of island configurations with and without pronouns, found that when extracting from an island, strong or weak, the resumptive structure was never judged "more grammatical" than its gapped version. Polinsky et al. (2013) proposed, then, that RPs do not establish an A' binding relationship, but a co-referential one. That is, RPs in English do not obey syntactic considerations but discourse-pragmatic ones, as they are considered anaphors. This was found in both on-line and off-line acceptability judgments. Likewise, McDaniel and Cowart (1999) found in an acceptability judgment task that native speakers of English did not prefer the resumptive pronoun over the trace structure in contexts in which the movement operation was illicit, i.e.: in island configurations, but they did prefer them in violations of conditions on representation. This made McDaniel and Cowart (1999) conclude that resumptive pronouns do not repair violations of the derivation (movement violations), and that they are spell-outs of traces. On the other hand, Ackerman et al. (2018), using several off-line forced-choice binary tasks, found that speakers of English strongly preferred RPs in island contexts, concluding that RPs indeed ameliorated island-violating sentences and questioned the assumed ungrammaticality of object-extracted resumptive pronouns in English.

More recently, in a further attempt to explain the syntactic and psycholinguistic nature of resumptive pronouns, Morgan and Wagers (2018) found a negative correlation between the acceptability of a gap structure and the production of RPs: as the acceptability of a structure with a gap decreases, the frequency of production of RPs in that structure increases. This result closely relates the production and comprehension domains, and indirectly rejects the idea that the production and the comprehension systems may consult different grammars, as Ferreira and Swets (2005) have suggested. Likewise, Chacón (2019) proposes that when speakers (comprehenders) try to resolve a filler-gap dependency, they do it preferably through a gap, which needs to be maintained in working memory over time. If working memory is strained though, then resumption becomes more acceptable. Thus, inasmuch as island configurations might suppose a burden for working memory, then they are a good host for resumption. To sum up, as this condensed review of studies dealing with resumptive pronouns in island configurations has shown, the paradox over RPs, —why are they produced by native speakers who rate them as unacceptable? —, as well as their nature —are they a processing entity or a syntactic one? —, are still open questions in the field, and even more so in SLA.

The general purpose of this study is to describe the nature of the Spanish interlanguage grammar of English and Arabic speaking learners by exploring *wh*-movement knowledge and its constraints. Ultimately, we want to determine whether L2 learners' knowledge is different or similar to that of a native speaker. With this in mind, we collected written production data of prepositional relative clauses as well as online grammaticality judgments on extractions from island configurations, in both conditions, with a gap or a resumptive pronoun. In turn, the acceptability data from our control group, the native speakers' data, will also serve us to reflect on the supposed (un)grammaticality of certain constructions, on the components that configure a grammatical judgment, and more in particular on the theory of *wh*-movement in L1 and L2. The following paragraphs will be devoted to explaining the three different strategies that prepositional relative clauses present in (Moroccan) Arabic, English and Spanish. The three possible syntactic strategies are Pied-Piping, Preposition Stranding, and Resumption.

Arabic, English and Spanish oblique relative clauses can be formed through Pied-Piping, a strategy which consists of moving the obligatory preposition along with the relative pronoun, as in (3). This strategy clearly involves *wh*-movement:

(3) Pied-Piping strategy:

a. L-katab	*'la-aš*$_i$	hd̲ar-ti t_i/
the-book	about-what	talked-you/

l-weld 'la-men$_i$ hd̲ar-ti t_i. *Moroccan Arabic*[1]
The boy about-whom talked-you

b. El libro *del* *cual*$_i$ hablaste t_i/
the book about-the which speak-you-past/
El chico *de quien*$_i$ hablaste t_i.
The book about who(m) speak-you-past

b' El libro/chico *de(l) que*$_i$ hablaste t_i. *Spanish*
the book/boy about-the that speak-you-past

c. The book *about which*$_i$ you talked t_i/
The boy *about whom*$_i$ you talked t_i. *English*

Moreover, English can leave the preposition dangling in its original position once the displaced constituent has moved; this option is ungrammatical in Spanish and Arabic, as the examples in (4) show, and involves movement:

(4) Preposition Stranding strategy:

a. The book (*which/that)*$_I$ you talked about t_i. *English*
a' The boy *who*$_i$ you talked about t_i.

b. *L-katab *aš/lli*$_i$ hd̲ar-ti *'la* t_i. *Moroccan Arabic*
the-book what/that talked-you about

c. *El libro *el cual/(el) que*$_i$ hablaste de t_i. *Spanish*
the-book the-which/(the)-that talked-you about

Finally, only Arabic accepts relative clauses with resumptive pronouns in its standard varieties. In fact, it is the most common strategy in standard Arabic, whereas it is ungrammatical or non-standard in English and Spanish, as the contrasts in (5) illustrate. This option in Arabic is not a last-resort strategy, as it could be the case in English or Spanish. In any case, resumptive pronouns appear always with complementizers and not with relative pronouns, as the contrasts among languages in (5) show.

(5) Resumptive pronoun strategy:

a. *The book which you talked about it./
*The boy who you talked about him. *English*
a' ??The book/boy (that) you talked about it/him.

b. L-katab *lli* hd̲ar-ti *'li-*(h)*./
the-book that talked-you about it/
L-weld *lli* hd̲ar-ti *'li-*(h)*. *Moroccan Arabic*
The-boy that talked-you about him

c' *L-katab *aš* hd̲ar-ti *'li-h*.
The-book what talked-you about it

c. *El libro *cual* hablaste *de él.*/
the book which speak-you-past about it/
*El chico *quien* hablaste *de él.* *Spanish*
The boy who speak-you-past about him

c' ?El libro *que* hablaste *deél.*
The book that speak-you-past about it

[1] Moroccan Arabic examples come or are adapted from Ennaji (1985).

The question that arises here is whether these Arabic resumptive constructions involve movement or binding. The standard analysis for Arabic is that resumption involves binding, and relativization of an argument out of an island configuration does not produce ungrammaticality, as illustrated with Lebanese Arabic in (6a). However, relativization of an adjunct is ungrammatical (6b) and this indicates that there was a violation of subjacency. These data seem to indicate that movement is available at least in some Arabic relative clauses. Aoun and Benmamoun (1998) presented further evidence from reconstruction effects that also points to a movement analysis for some Arabic relatives. Examples from Aoun and Benmamoun (1998):

(6) a. mnaʕrif l-mara lli fallayto ʔabl
Know.1p the-woman that left.2p before
ma yʔeebəl-**a** Karim
Comp meet.3sm-her Karim
'we know the woman that you left before Karim met (her)'

b. *ssərʕa lli btaʕrfo miin byiʃtiɣil fiy-*(a)
the-speec that know.2p who works with-(*it)
hiyye l-maṭluube
she the-required
'The speed with which you who works is the required one'

As for Spanish, Suñer (1998) proposed that it is a language that has two types of resumptive pronouns, those optionally inserted in all types of relative clauses (direct and indirect object, prepositional, subject, genitives, locatives), at the level of PF, and those obligatory, used as a last resort, to prevent the structure from an island violation. This type of last resort resumptive pronoun exists in Spanish (7a) and in English (7b), and is the focus of our investigation.

(7) a. ¿Qué [libro]$_i$ me dijiste que no recuerdas
Which book to.me you-told that not you-remember
[dónde]$_j$ *(LO)$_i$ pusiste t_i t_j?
where IT you-put
'Which book did you tell me that you don't remember where you put (it)?

b. The settlement that Caroline asked [when] we would get *(IT) (Suñer, 1998:335)

The specific purpose of this study is to first investigate the availability of *wh*-movement in prepositional relative clauses in L2 Spanish, and second, to investigate the grammatical nature of gapped and resumptive islands in L2 learners whose native languages present both *wh*-movement (English and Arabic) and resumptive pronouns (Arabic). Ultimately, we aim to reflect on the concept of grammaticality through acceptability ratings in both native and interlanguage grammars, the reliability of

experimental and introspective data, and how these have been used to argue for or against L2 learners' accessibility to UG.

WH-MOVEMENT AND SUBJANCENCY IN L2 LEARNERS

The availability of *wh*-movement has been the central issue of many studies that discussed accessibility to UG and the differences and similarities between L1 and L2 acquisition (Johnson and Newport, 1991; Hawkins and Chan, 1997; White and Juffs, 1998; among others). In the late 80s, subjancency violations were one of the main arguments for the Fundamental Difference Hypothesis in Second Language Acquisition (Bley-Vroman, 1990; Johnson and Newport, 1991). The early L2 studies on subjacency violations mostly included learners whose L1 does not present overt *wh*-movement, such as Korean or Chinese (Bley-Vroman et al., 1988; Schachter, 1990; Johnson and Newport, 1991; White and Juffs, 1998). For instance, Chinese is a language that does not present overt *wh*-movement, at least with argumental *wh*-movement (Huang, 1982). Johnson and Newport (1991), and Hawkins and Chan (1997) found that the Chinese-speaking learners had problems recognizing subjacency violations in English, a result that made these researchers argue that L2 learners do not have full access to UG, otherwise they would respect the universal principle of subjacency. On the other hand, White and Juffs (1998) found that Chinese speakers with more advanced knowledge of English were accurate at judging these violations, arguing that these L2 learners could indeed access UG. Another general finding in these studies that was later noticed is that performance significantly varied depending on the type of island configuration, L2 learners rejecting strong islands (relative clauses and subjects) more accurately than weak islands (*wh*-islands and noun complements) (Martohardjono, 1993). That is, L2 learners perceived the gradience in grammaticality, which Schwartz and Sprouse (2000) interpreted as an indication of UG access since none of these types of island configurations, weak and strong, are present in the input. This grammaticality asymmetry was accounted for in the revised CED (Huang, 1982; Nunes and Uriagereka, 2000), in which it is stated that subjects and adjuncts are universally islands, as opposed to *wh*-islands, which might be parameterized. Therefore, if L2 learners are not consistent at rejecting weak islands such as *wh*-islands, then these data cannot really inform us about the L2 learners' accessibility to UG. This is one of the main points raised by Belikova and White (2009), which concluded that, even though islands still constitute a typical poverty of the stimulus scenario, these are now understood to be regulated by computational principles in all languages, and thus, do not speak toward the accessibility to UG, or the difference between L1 and L2 acquisition. The present study reinforces these general conclusions and further questions the assumed grammaticality of certain island configurations.

More recently, L2 studies have implemented on-line methodologies to assess the real-time processing of *wh*-dependencies and island constraints, and to investigate whether L2 learners are able to use syntactic information in real-time processing (Aldwayan et al., 2010; Omaki and Schulz, 2011; Kim et al., 2015; Johnson et al., 2016 a. o.). For instance, Aldwayan et al. (2010) investigated whether Najdi Arabic (a *wh-in situ* language with obligatory resumption) learners of English have the knowledge of syntactic constraints in the processing of *wh*-movement and whether they process these structures incrementally. With a self-paced reading task, they showed that advanced L2 learners are guided by syntactic constraints and posit gaps during incremental language processing, as native speakers do, disproving the Shallow Structure Hypothesis (Clahsen and Felser, 2006). Similarly, Aldosari (2015) found that Najdi Arabic speakers who are learners of English were sensitive to syntactic island constraints on *wh*-movement, and that individual differences such as working-memory capacity did not have an effect on sensitivity to island effects, concluding that islands are not due to limited processing resources but most likely to syntactic constraints. With respect to Spanish-speaking learners of English, Kim et al. (2015) found that Spanish speakers did not keep active a filler-gap dependency in a relative clause island configuration, obeying the same restrictions as native speakers. These authors did not exactly find the same results in Korean learners of English (Korean being a *wh-in situ* language), who seemed to have posited a gap when processing an island configuration even though they showed knowledge of *wh*-movement restrictions in islands in the off-line task. Kim et al. (2015) interpreted these results by proposing that the L1 of the learners influences the L2 learners' processing. None of these studies, though, directly tackled the issue of resumptive pronouns in SLA, the focus of our investigation.

In order to assess whether L2 learners know the limits of *wh*-movement and the locality constraints that regulate it, first it must be determined that the learners indeed have *wh*-movement in their interlanguage grammars. Some of these studies included *wh*-question formation to show that movement was already mastered, but there is some controversy with this procedure since *wh*-questions can imply topicalization or scrambling, in which movement is not involved. For these reasons, we decided to include relative clause formation in our study. As shown in (3) above, all languages at play in this study can form oblique relative clauses through movement (Pied-Piping); English can also employ Preposition Stranding, another movement structure, and Arabic usually resorts to resumptive pronouns in its relative clauses, a no-movement option. In this study we want to investigate the limits of *wh*-movement in L2 learners whose native language already presents *wh*-movement, an understudied combination. It has typically been the case in the literature that problems rejecting island violations were explained by the lack of *wh*-movement in the L1s of the L2 learners. However, it has not been investigated whether those grammaticality judgments assigned to island configurations were a true reflection of the inability to constrain *wh*-movement, or whether these were measuring a different type of linguistic phenomenon in the L2 learners' interlanguage. It could be the case that comprehension of island configurations goes beyond the realm of *wh*-movement. This is what we aim to unravel in this study.

Related to the (in)ability to displace *wh*-elements and to create filler-gap dependencies, we also included islands rescued

by resumptive pronouns. Resumptive islands in SLA have been hardly investigated, not even in L2 learners whose native language accepts resumptive pronouns in relative clauses, such as the case of Arabic. We believe that, if we want to investigate the nature of island configurations and more particularly the nature of the grammaticality judgments of island configurations in both L1 and L2, resumptive islands need to be included in the experimental design, particularly if one of the languages at play presents resumptive pronouns in its standard variety. Thus, this study is twofold: by focusing on the properties of *wh*-movement in interlanguage grammars and questioning some of the commonly accepted assumptions for island configurations, it aims to generally reflect on the concept of grammaticality in SLA and the theoretical hypotheses that hinge on it. This study has three general research questions (RQ1a, b, c) and two specific research questions (RQ2a, b):

RQ1. a. What do grammaticality judgments tell us about the nature of interlanguage grammars and the native knowledge of a language?
b. What do judgments on island configurations tell us about *wh*-movement theory?
c. What do judgments on island configurations tell us about the (in)ability to *wh*-movement in a second language grammar?

RQ2. a. Would L2 learners whose native language already presents some type of *wh*-movement strategy also employ *wh*-movement when forming oblique relative clauses in an L2?
b. Would L2 learners be able to constrain *wh*-movement appropriately in their second language by rejecting island violations and accepting resumptive islands?

Considering the linguistic phenomenon under investigation and its properties in English and Arabic described in (3–5), we formulate the following hypothesis for the specific research questions:

- H1: Assuming the Full Transfer/Full Access Hypothesis (FT/FAH, Schwartz and Sprouse, 1994, 1996), which postulates full transfer of the L1 and full access to UG in L2 acquisition, if the L1 is fully transferred into the L2 grammar, then, the L2 learners would be able to employ *wh*-movement when forming oblique relative clauses in L2 Spanish. That is, we should not expect L2 learners to have major problems constructing oblique relative clauses through Pied-Piping because this strategy is already present in their L1s.
- H2: Also, assuming the FT/FAH, we could expect some degree of negative transfer, such as Preposition Stranding in English L2 learners' grammars, and Resumption in Arabic L2 learners' grammars, especially at early stages of development.
- H3: Finally, if participants already have *wh*-movement in their L1s, then we will find that relative clauses formed as an extraction from an island will be judged as ungrammatical due to subjacency violations. If, on the other hand, they interpret relative clauses through binding and not movement, then these participants will accept ungrammatical extractions out of an island. In both cases, we expect participants to accept extractions from islands rescued by a resumptive pronoun.

THE STUDY

In order to investigate these questions on the nature of interlanguage and native grammars and *wh*-movement knowledge, we designed a series of tasks. Here, we are reporting the results of two of these tasks: a written production task that elicited relative clauses, and a self-paced grammaticality judgment task with different types of island configurations. The data we are analyzing in this study is part of a series of experiments on the L2 processing and L2 acquisition of relative clauses (Perpiñán, 2010).

Participants

An initial pool of 20 native Spanish speakers and 116 Spanish learners (L1 English or L1 Arabic) participated in this study. The English-speaking learners (n = 81) were college students enrolled at the University of Illinois or at the Knox College at the time of testing (mean age = 21.9). They were all born and raised in the United States, and they were recruited either at intermediate or advanced Spanish courses. Students who used a different language at home (Korean, Polish, Spanish, etc.) and who knew other second languages (as reported on the background questionnaire) were excluded from the data analysis. The Arabic speakers (n = 35) were all native speakers of the colloquial Moroccan Arabic variety or "dariʒa". Native speakers of other languages such as Berber were excluded from the experiment. The Arabic speakers were students of intermediate or advanced Spanish courses either at the Instituto Cervantes or at the language academy "Dar Loughat" in Tetouan, Morocco. Most of them were college students although there were also some civil servants or professionals in the pool (mean age = 25.6). Since it is impossible to find educated participants in Morocco, who have not studied French or have taken courses in French, these subjects are, potentially, L3 speakers of Spanish. However, most of them reported that their knowledge of French was limited and that they felt more comfortable speaking in Spanish than they did in French. The control group consisted of native speakers of Spanish (n = 20), 8 males and 12 females, from different dialectal varieties: one Argentinean, one Colombian, one Costa Rican, one Mexican, one Venezuelan, and fifteen speakers of Castillian Spanish. Their mean age at the time of testing was 32.25. All but two were college graduates.

All participants took a proficiency test, which consisted of a slightly modified version of the standardized grammar section of the superior level of the Diploma de Español como Lengua Extranjera (DELE), created by the Instituto Cervantes. In this proficiency test we included six screening items that tested subcategorization knowledge of the prepositional experimental verbs: *hablar de* (to talk about), *depender de* (to depend on), *pensar en* (to think about), *confiar en* (to rely on), *soñar con* (to dream about), *contar con* (to count on). These verbs required a

preposition in the three languages we are considering: Spanish, English and Moroccan Arabic. Participants who did not know that these verbs subcategorized a prepositional argument were not invited to continue with the study. After this scrutiny, only 42 L2 learners (21 English speakers/21 Arabic speakers) completed the entirety of the experiment. The participants' proficiency scores (maximum score 40) were submitted to a one-way ANOVA, and as expected, the results of the ANOVA indicated a significant effect by group $F(2,59) = 28.74$, $p < 0.001$. A *post hoc* Tukey HSD test revealed that the only different group was the control group ($p < 0.001$), whose mean score was 39.6 (SD.681), with a 99% rate of accuracy. The Arabic (mean score = 25.67, SD = 8.79, 64% accuracy) and English learners of Spanish (mean score = 26.05, SD = 7.32, 65% accuracy) did not differ significantly ($p = 0.98$).

TASK 1: WRITTEN PRODUCTION TASK

The purpose of this task was to reveal how productive our participants' *wh*-movement structures are. Participants were presented with two independent sentences that shared one constituent and were instructed to combine the two sentences, retaining the same meaning while not using the repeated constituent again. The beginning of each new sentence was provided to ensure that the participants used that constituent as the extracted part of the complex sentence. Two examples were provided: the first one demonstrated a prepositional construction and thus, a Pied-Piped relative clause; the second exemplified a transitive construction. The experiment included the 6 target items that required prepositional RCs and 5 items targeting direct object RCs. In this study, we are only interested in the prepositional contexts. Examples are shown in (8) below.

(8) Examples provided in written sentence-combining task:
a. El parque es muy bonito. Cada tarde iba
a ese parque.
El parque *al que iba cada tarde es muy bonito.*

'The park is very nice.
Each afternoon I/She-went to that park.
The park to which I/(S)he-went each afternoon
is very nice.'

b. Esa canción es mi preferida. Juan cantó esa canción.
La canción *que cantó Juan es mi preferida.*

'This song is my favorite. Juan sang that song.
The song that Juan sang is my favorite.'

RESULTS TASK 1: WRITTEN PRODUCTION OF RELATIVE CLAUSES

A total of 682 sentences were generated in the written experiment; 372 in the prepositional context are the only ones that we will consider here (see Perpiñán, 2013 for more data). Sentences were coded according to their structure, and frequencies and raw numbers (in parentheses) are calculated for each structure produced; data are displayed in **Table 1**. In order to compute non-parametric statistics on these categorical data, sentences were coded as "target-like" vs. "non-target-like." Hence, the baseline for comparison is not the native speakers' production but the expected construction for each group.

Out of the 372 sentences produced, only 257 were target-like. Native speakers behaved as expected, and 99.2% of their sentences were formed through Pied-Piping, but only 62.7% of the English learners' production and 46.8% of the sentences produced by the Arabic learners were target-like, that is, formed through Pied-Piping.

The percentages alone already seem to indicate that there is a significant difference among the three groups, as the Chi square based on the accuracy of the sentences × groups demonstrates χ^2 (2) = 82.48, $p < 0.001$. Furthermore, the two experimental groups (English vs. Arabic speakers) also differed significantly χ^2 (1) = 6.407, $p = 0.011$ between themselves, as English speakers were more target-like than the Arabic speakers. And since the native group only missed one sentence out of 120, the odd ratios are enormous: English speakers were 70.8 times more likely to be non-target-like than the native group, and in the case of the Arabic speakers, the inaccuracy ratio compared to the control group is up to 135.2. Thus, although Spanish prepositional relative clauses present some difficulties for L2 learners, the target Pied-Piping is nonetheless the most produced construction in both groups.

The deviance from the target structure by the English-speaking learners not only consisted of producing the ungrammatical L1 transferred structure Preposition Stranding, as in (9a), but also a relative clause without the obligatory preposition, a phenomenon termed Null Prep by Klein (1993), such as (9b). The same holds for the Arabic speakers who produced 22.2% of these sentences without the obligatory preposition, as in (10a), and 20.6% of the sentences with the preposition and a strong resumptive pronoun, as in (10b). All instances of RPs appeared with the complementizer "que."

(9) a. La amiga quien María confiaba
'The friend who María relied
en es una mentirosa. (L2 Engl. # 20)
on is a liar.'

b. El hombre Ø que María depende económicamente
'The man that María depends economically
es muy rico. (L2 Eng. # 13)
is very rich.'

(10) a. La chica Ø que mis amigos hablan frecuentemente
'The girl that my friends talk_3p frequently
es muy guapa. (L2 Ar. # 45)
is very beautiful.'

b. La muchacha que Juan pensaba **en ella** a todas horas
'The girl that Juan thought about her at all hours
es guapísima. (L2 Ar. # 30)
is very-beautiful.'

TABLE 1 | Frequency of constructions produced in written prepositional RC, percentages and raw numbers.

Group	Pied-piping	Null prep	Preposition stranding	Resumptive	No RC	Other	Total
Natives	99.2 (119)	0	0	0.8 (1)	0	0	100 (120)
L2 English	62.7 (79)	15.9 (20)	16.7 (21)	0	3.2 (4)	1.6 (2)	100 (126)
L2 Arabic	46.8 (59)	22.2 (28)	0	20.6 (26)	5.6 (7)	4.8 (6)	100 (126)

TASK 2: SELF-PACED GRAMMATICALITY JUDGMENT TASK

Procedure

The self-paced reading task consisted of a total of 84 items followed by a yes/no grammaticality judgment question. Half of the sentences were grammatical, and half ungrammatical. 24 of these sentences were relative clauses (see Perpiñán, 2015), 18 items tested subjacency constraints, our experimental conditions, and the remaining 42 sentences were distracters. Sentences were pseudorandomized so that no token from the same condition would appear consecutively. Participants (the same ones as in the previous task) had to read the sentences in a self-paced, non-cumulative word-by-word display on a computer monitor, using the experimental software Linger. The segments initially appeared as a row of dashes, and participants pressed the space bar on the keyboard to reveal each subsequent word of the sentence. At the end of each sentence, participants had to answer the question "Esta frase, ¿está bien?," *('This sentence, is it ok?')* and then answer as quickly as possible pressing the keys "F" for *yes* and "J" for *no*. These keys were shown in a different color on the keyboard. Participants received immediate feedback if they responded differently than expected: "¡Oh, lo siento!" (Oops, I'm sorry). This feedback was mainly included to encourage participants to stay focused on what they were reading. Nevertheless, all participants were instructed to follow their intuition when judging the sentences, regardless of the feedback prompted. In fact, they were warned that the computer was not always right and that it was legitimate not to agree with the computer's feedback.

Stimuli

The results of the written production task served us to select the three types of extraction from an island that we included in the GJT: Pied-Piping extraction, Null-Prep extraction, and extraction with a resumptive pronoun in the island configuration. We chose strong islands (if-clauses) since previous literature has shown that weak islands might be parameterized and do not hold in all languages, and that L2 learners are mostly sensitive only to this type of islands. Participants needed to make a judgment about the grammaticality of the sentence as fast as possible. The head of the relative clause was extracted from a strong island, specifically a conditional clause. The relative clause was formed either through Pied-Piping, Null-Prep or Resumption. There were six items per condition, one item per each experimental prepositional verb (*depender, hablar, pensar, contar, soñar, confiar*) ($3 \times 6 = 18$ island-type sentences). To avoid confusion, the pseudorandomization ensured that no island sentence of any type would appear right after another island sentence. Also, and since these were long distance extractions, we made sure that the extracted constituent could not be interpreted as an argument of the antecedent of the conditional clause. For this reason, only intransitive verbs were included in this position such as *dormir* ("to sleep"), *callar* ("to shut up"), or *respirar* ("to breath").

The control structure was the Pied-Piping island configuration (11). There is no disagreement with respect to the ungrammaticality of this construction since Pied-Piped relative clauses undoubtedly involve wh-movement.

(11) Pied-Piping Island Configuration
*El hombre$_i$ en el que Marta sería feliz t_i si Pedro no
The man on the that Marta be.COND happy if Pedro not
pensara t_i continuamente es muy alegre.
think.PAST.SUB continuously is very cheerful.
'The man of whom Marta would be happy if Pedro didn't think continuously is very cheerful.'

- Question prompted: *Esta frase, ¿está bien?* Expected response N.

On the other hand, it is generally assumed that resumptive relative clauses do not engage movement and are interpreted through A-bar binding. For this reason, resumptive island configurations were coded as grammatical (12). In fact, the appearance of resumptive pronouns in island configurations is typically described as a last resort mechanism to rescue the derivation from the ungrammaticality.

(12) Resumptive Island Configuration
La mujer$_i$ que Juan respiraría mejor
The woman that Juan breath.COND_cond. better
si Pedro no soñara frecuentemente **con**
if Pedro not dream.PAST.SUB frequently with
ella$_i$ es inteligente.
her is very intelligent
'The woman that Juan would breath better if Pedro did not dream about her frequently is very intelligent.'

- Question prompted: *Esta frase, ¿está bien?* Expected response Y.

Finally, we also included in the experiment an island configuration with a relative clause formed through the Null-Prep strategy. This strategy was significantly produced by all L2 learners, and for this reason, we have decided to include it. This island configuration was *a priori* coded as ungrammatical, as a relative clause formed through Null-Prep.

(13) Null-Prep Island Configuration
*La mujer$_i$ que Marcos dormiría mejor t_i si Pedro no
The woman that Marcos sleep.COND better if Pedro not
dependiera t_i económicamente es muy fuerte.
depend.PAST.SUB economically is very strong
'The woman that Marcos would sleep better if Pedro did not depend economically is very strong.'

- Question prompted: *Esta frase, ¿está bien?* Expected response N.

RESULTS TASK 2: SELF-PACED GTJ

Accuracy was measured in average proportions, from 0 to 1 depending on the expected answer, where 1 indicated that the response given matched the codification made for that condition (correct response), and 0 indicated that the response given did not match the expected response (incorrect response). However, in order to understand the results independently from the aprioristic coding, accuracy was transformed into acceptability. This way, acceptability computes whether the participants judged the sentences as ok ("está bien") or not ok ("no está bien") regardless of the expected response. In these measurements, 0 means that the participant thought that the sentence was not ok, (not accepted) whereas 1 means that the sentence was ok (accepted). The average of these responses was calculated per structure and person. **Figure 1** displays the acceptability averages per group and structure, with the Standard Error of the group.

The first interesting result is that native speakers overwhelmingly considered the sentences not ok, that is, ungrammatical. The L2 learners, on the other hand, do not seem to have robust intuitions regarding the acceptability of these sentences, accepting these sentences as adequate around 40–45% of the time. The proportions of acceptability responses were Arcsine transformed to account for their binomial distribution, and later submitted to a mixed-design repeated measures ANOVA with island structure (Pied-Piping, Null-Prep, Resumption) as a within-subjects factor, and group (native, L1 English, L1 Arabic) as a between-subjects factor. The Mauchly's test indicated that the assumption of sphericity was not violated ($\varepsilon = 0.991$), and the within-subjects results revealed a mild main effect for structure [$F(2, 120) = 3.706$, $p = 0.027$, $\eta_p^2 = 0.058$], a main effect of group [$F(2, 60) = 314.62$, $p < 0.001$, $\eta_p^2 = 0.328$], but no significant interaction between structure and group ($p > 0.05$). The *post hoc* test for group indicated that the native speakers' group was different from the two L2 learners' groups ($p < 0.001$), and the two experimental groups did not differ between them ($p = 1$). We further explored the differences in structure and found that Pied-Piping was overall different from Null-Prep [$F(1, 60) = 5.93$, $p = 0.018$, $\eta_p^2 = 0.090$], and from the Resumptive condition [$F(1, 60) = 4.61$, $p = 0.036$, $\eta_p^2 = 0.071$]. However, when we carried out the within-subjects analysis independently for each group, the tests revealed that the main effect for structure only held in the native speaker group [$F(2, 38) = 7.214$, $p = 0.002$, $\eta_p^2 = 0.275$], but not in the learners' groups ($p > 0.1$). Likewise, only the native group distinguished between the Pied-Piping island condition and the Null-Prep island condition [$F(1, 19) = 12.53$, $p = 0.002$, $\eta_p^2 = 0.397$], and between the Pied-Piping island and the Resumptive island [$F(1, 19) = 8.953$, $p = 0.007$, $\eta_p^2 = 0.320$]; all the other contrasts were not significant ($p < 0.01$). To summarize so far, only the native speakers distinguished among the different types of islands, in favor of the gapped island, which was generally judged as more acceptable than the other two, against what has been reported in the theoretical literature.

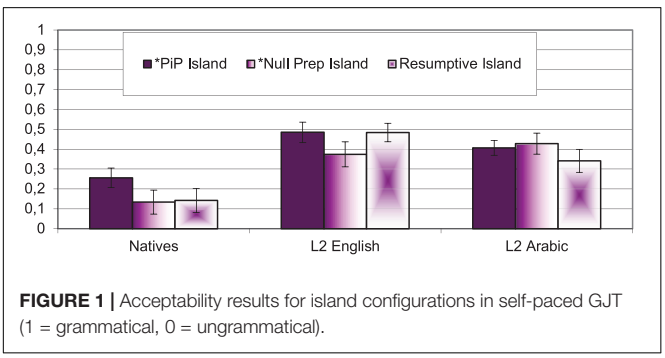


FIGURE 1 | Acceptability results for island configurations in self-paced GJT (1 = grammatical, 0 = ungrammatical).

DISCUSSION

In this study, we want to reflect on the (un)acceptability of island configurations in both L1 and L2, and its relation to the availability to *wh*-movement in these grammars. First, we will discuss the unexpected results from the native speakers and what these could mean for linguistic theory, and, in particular, for the theory of *wh*-movement, taking into account some psycholinguistic considerations. Later, we will discuss the data of the L2 learners and their implications for our views on the nature of interlanguage grammars.

The first main finding of this study is that native speakers, our control group, do not distinguish among island violations, and crucially, the resumptive pronoun does not improve the acceptance rates of these sentences. This is at odds with the traditional literature on island configurations and particularly with the assumed rescue effects of resumptive pronouns. Nevertheless, similar findings have been attested in McDaniel and Cowart (1999) with a relative acceptability judgment task for English relative clauses and islands; in Heestand et al. (2011) and Polinsky et al. (2013), studies devoted to the off-line and online comprehension of gapped and resumptive island constructions in native speakers of English, and in Alexopoulou and Keller (2007, 2013). In all of these experimental studies, it was found that when extracting from an island, the ungrammatical gapped condition was judged equal if not more acceptable than the supposedly "rescued" version with a resumptive pronoun. Our study corroborates these findings additionally for Spanish, as our native speakers found all extractions from island configurations unacceptable, both with a gap or a resumptive pronoun. Indeed, Spanish native speakers more often accepted the extraction with Pied-Piping from an island, which involves illicit *wh*-movement, than extractions from islands repaired with a resumptive

pronoun. This is a novel result as, to our knowledge, Pied-Piping island configurations were not tested before, in English or in Spanish. It could be the case that the complexity of the extracted element (P+ *wh*-word) makes it more salient and/or more referential, and as such, it remains highly activated in memory (Just and Carpenter, 1992; Kluender, 1998; Hofmeister and Sag, 2010), making its integration in the discourse (d-linking) easier. These data would corroborate the main ideas of Hofmeister and Sag (2010) who propose that the unacceptability of island configurations goes beyond their syntactic nature, and is (also) motivated by the interaction of other cognitive constraints such as referentiality, saliency, d-linking, and/or the complexity of the filler phrase.

Granted, island sentences are difficult to judge, and require certain training and time, which the participants did not have. One of the reasons for choosing a timed GJT was to get the first, less conscious intuition about the structure. This would go with a generative view of language, which considers that real time construction of grammar sometimes loses grammar accuracy (Chomsky and Lasnik, 1993; Townsend and Bever, 2001), and against a view in which real-time processing can capture fine-grained distinctions (Phillips, 2003, 2006). In fact, this is not the only experiment which has failed to discover island sensitivity in processing experiments. For instance, Frazier and Clifton (1989) showed acceptance of gaps inside an island using speeded grammaticality tasks. Ferreira and Swets (2005) found dissociation between the production system and the comprehension system with respect to resumptive pronouns in island contexts. They found that native speakers of English judged these sentences as unacceptable in the grammaticality judgment task, but at the same time, they produced resumptives in islands in an oral experiment. Moreover, Ferreira and Swets (2005) further concluded that the "marginal" structure (resumptive island) takes more processing resources to produce, and participants found them harder to understand than a similar but grammatical construction. In their oral production experiment, the resumptive island construction was more often produced in the no time pressure condition than in the time-constrained condition, a result that the authors interpret as a sign of its costly nature, particularly with a RP. On the other hand, Chacón (2019) relates the appearance of RPs with long filler-gap dependencies that strain on working memory resources. That is, the RP appears as an anaphoric way to resolve the filler-gap dependency when the representation of the gap has failed. Similarly, Morgan and Wagers (2018) found that the production of RPs increases as the acceptability of a gap decreases. In any case, these proposals relate RPs with processing costs, implying that island configurations are not only a syntactic entity. This is also the position we take here. What seems to be clear from the experimental data gathered from GJTs is that RPs do not ameliorate island configurations; likewise, in this study, we failed to find an acceptability improvement of islands "repaired" by RPs, even in speakers who still produce RPs in relative clause formation, and whose native language (Arabic) accepts and requires RPs in these contexts. We interpret these results as a clear indication that islands, with gaps or with RPs, are not a purely syntactic phenomenon, and that using them as a means to determine the accessibility of L2 learners to UG is a moot point, as Belikova and White (2009) already concluded.

It must be acknowledged that the sentences included in the present experiment do not make complete sense, regardless of their grammatical status. In other words, these sentences are experimental in nature and are quite implausible, and we know that plausibility is a very relevant factor when interpreting sentences in real time (Traxler and Pickering, 1996; Pickering and Traxler, 1998; Pickering et al., 2000). Besides, there are several studies that have found that self-embedded sentences, such as the ones used in this experiment, are very hard to process due to memory capacity. This is the case because the reader needs to hold what has been read in memory for a long time, while also integrating new entities into the discourse (Lewis, 1996). Consequently, non-local dependencies are usually problematic not only for L2 learners (Dallas and Kaan, 2008) but also for monolingual native speakers (Gibson, 1998). The processing load of reading, memorizing and integrating meaning on-line makes comprehension and grammaticality judgments more difficult than in untimed tests. In the on-line GJT, there are factors such as word segmentation, memory or disruptions that play a significant role in quick decision making. The fact that paper and pencil experiments have found similar results indicates that all these factors are relevant and active when processing island constraints under no time pressure. Due to all this, we believe that island interpretation is a multifactorial matter, and that to isolate the most significant factors that contribute to their interpretability is very difficult, if not impossible. For instance, Kluender (1998) proposed that it is the interaction between verbal working memory and referential processing that explains the traditional dichotomy between strong and weak islands, and that, in the end, "wh-islands are essentially an interpretive problem" (Kluender, 1998:243).

These same considerations apply to the L2 learners' processing, whose results are even less conclusive than those from the native speakers. Firstly, the production data indicates that for the most part, our L2 speakers form relative clauses through movement, particularly the English-speaking group. As for the Arabic group, 20.6% of their relative clauses are formed with a resumptive pronoun, and only three speakers constructed all relative clauses with the resumptive strategy, that is, without *wh*-movement, as hypothesized in H2. Assuming that Null-Prep relative clauses are also formed through movement, we can conclude that our L2 learners (except for those three Arabic speakers) know the rudiments of *wh*-movement in relative clauses, as hypothesized in H1. Still, they have very weak intuitions about the grammaticality of extractions from island configurations, and they tend to accept these (un)grammatical sentences between 40–50% of the time. Likewise, the L2 learners do not distinguish among the three types of extractions from islands, and similarly to the native speakers, do not have a preference for resumptive islands, that is, RPs do not improve their judgments about islands. One possible explanation for these results is to pose that native and L2 speakers alike tried to interpret resumptive islands through movement, as it would be the case with any other extraction. It is only after a processing failure that these sentences are interpreted through binding,

and the RP is not able to repair the processing failure at this point. We favor an explanation — not incompatible with the previous one — which does not necessarily take these judgment data at face value. That is, it does not automatically condemn these resumptive structures and proposes that the speakers might not be judging the grammaticality of the sentence, but the plausibility, the naturalness, the depth of embedding, or simply that what we are measuring is the processability of this long sentence, and not its grammatical well-formedness.

Where do these data leave us in terms of the appropriateness of the methodology for our research purposes? How can we measure L2 knowledge of a phenomenon for which the native language does not provide a clear baseline? Crucially, our L2 learners, despite their weak intuitions, do not present the assumed contrast between gapped and resumptive islands, not even the learners whose native language presents resumptive pronouns in standard relative clauses (Arabic); but neither do native speakers. We suppose, then, that L2 learners are sensitive to the same type of processing and interpretative factors that native speakers are, even when their knowledge might still be in progress and present transfer effects, as found in Perpiñán (2015). This means that the L2 learners' — and probably also the native speakers' – processing might be somewhat dissociated from their grammatical knowledge, and even though the L2 grammatical representation might not be fully complete, the learners are able to grasp some of the interpretative and processing factors that condition the grammatical judgments on island configurations. In light of these results, this study contributes to the line of reasoning opened by Belikova and White (2009) and casts doubt on the suitability of assessing accessibility to UG by testing *wh*-islands, as it was typically done during the 90s. That is, if *wh*-islands are not a purely representational issue but an epiphenomenal one whose acceptability goes beyond grammatical well-formedness, for both native and non-native speakers alike, then GJTs on islands are not a reliable way to assess L2 grammatical knowledge. Still, they give us precious information on the way speakers interpret these sentences and whether L2 learners and native speakers resort to the same mechanisms while processing complex sentences.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the IRB, University of Illinois, Urbana-Champaign. Protocol Number: 08330. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Ackerman, L., Frazier, M., and Yoshida, M. (2018). Resumptive pronouns can ameliorate illicit island extractions. *Linguist. Inq.* 49, 847–859. doi: 10.1162/ling_a_00291

Aldosari, S. (2015). *The Role of Individual Differences in the Acceptability of Island Violations in Native and Non-Native Speakers*. Ph.D. dissertation, University of Kansas, Lawrence, KS.

Aldwayan, S., Fiorentino, R., and Gabriele, A. (2010). "Evidence of syntactic constraints in the processing of wh-movement: a study of Najdi Arabic learners of English," in *Research in Second Language Processing and Parsing*, eds B. VanPatten, and J. Jegerski (Amsterdam: John Benjamins), 65–86. doi: 10.1075/lald.53.03ald

Alexopoulou, T., and Keller, F. (2007). Locality, cyclicity, and resumption: at the interface between the grammar and the human sentence processor. *Language* 83, 110–160. doi: 10.1353/lan.2007.0001

Alexopoulou, T., and Keller, F. (2013). "What vs. who and which: kind-denoting fillers and the complexity of whether-islands," in *Experimental Syntax and Island Effects*, eds J. Sprouse, and N. Hornstein (Cambridge: Cambridge University Press), 310–340. doi: 10.1017/cbo9781139035309.016

Aoun, J., and Benmamoun, E. (1998). Minimality, reconstruction, and PF movement. *Linguist. Inq.* 29, 569–597. doi: 10.1162/002438998553888

Aoun, J., Choueiri, L., and Hornstein, N. (2001). Resumption, movement, and derivational economy. *Linguist. Inq.* 32, 371–403. doi: 10.1162/002438901750372504

Asudeh, A. (2012). *The Logic of Pronominal Resumption*. Oxford: Oxford University Press.

Belikova, A., and White, L. (2009). Evidence for the fundamental difference hypothesis or not? Island constraints revisited. *Stud. Second Lang. Acquis.* 31, 199–223. doi: 10.1017/s0272263109090287

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguist. Anal.* 20, 3–49.

Bley-Vroman, R. (2009). The evolving context of the fundamental difference hypothesis. *Stud. Second Lang. Acquis.* 31, 175–198. doi: 10.1017/s0272263109090275

Bley-Vroman, R. W., Felix, S. W., and Ioup, G. L. (1988). The accessibility of universal grammar in adult language learning. *Second Lang. Res.* 4, 1–32. doi: 10.1177/026765838800400101

Chacón, D. A. (2019). "How to make a pronoun resumptive," in *Proceedings of the 36th West Coast Conference on Formal Linguistics*, eds R. Stockwell, M. O'Leary, Z. Xu, and Z. L. Zhour (Somerville, MA: Cascadilla Proceedings Project), 99–108.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1973). "Conditions on transformations," in *A Festschrift for Morris Halle*, eds S. R. Anderson, and P. Kiparsky (New York, NY: Holt, Rinehart & Winston), 232–286.

Chomsky, N. (1977). "On WH-movement," in *Formal Syntax*, eds P. Culicover, T. Wasow, and A. Akmajian (Cambridge, MA: Academic Press), 71–132.

Chomsky, N. (1981). *Lectures on Government and Binding, The Pisa Lectures*. Berlin: De Gruyter Mouton.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Chomsky, N., and Lasnik, H. (1993). "The theory of principles and parameters," in *An International Handbook of Contemporary Research*, eds J. Jacobs, A. von Stechow, W. Sternefeld, and T. Vennemann (Berlin: Walter de Gruyter).

Clahsen, H., and Felser, C. (2006). Grammatical processing in language learners. *Appl. Psycholinguist.* 27, 3–42. doi: 10.1017/s0142716406060024

Dallas, A., and Kaan, E. (2008). Second language processing of filler-gap dependencies by late learners. *Lang. Linguist. Compass* 2, 372–388. doi: 10.1007/s10936-009-9104-8

Ennaji, M. (1985). *Contrastive Syntax: English, Moroccan Arabic, and Berber Complex Sentences*. Würzburg: Königshausen & Neumann.

Ferreira, F., and Swets, B. (2005). "The productions and comprehension of resumptive pronouns in relative clause "island" contexts," in *Twenty-First Century Psycholinguistics: Four Cornerstones*, ed. A. Cutler (New York, NY: Lawrence Erlbaum Associates), 263–278.

Frazier, L., and Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Lang. Cogn. Process.* 4, 93–126. doi: 10.1080/01690968908406359

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/s0010-0277(98)00034-1

Hawkins, R., and Chan, C. Y. (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

Heestand, D., Xiang, M., and Polinsky, M. (2011). Resumption still does not rescue islands. *Linguist. Inq.* 42, 138–162.

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2013). "Islands in the grammar? Standards of evidence," in *Experimental Syntax and Island Effects*, eds J. Sprouse, and N. Hornstein (Cambridge: Cambridge University Press), 42–63. doi: 10.1017/cbo9781139035309.004

Hofmeister, P., and Sag, I. A. (2010). Cognitive constraints and island effects. *Language* 86, 366–415. doi: 10.1353/lan.0.0223

Hopp, H. (2007). *Ultimate Attainment at the Interfaces in Second Language Acquisition: Grammar and Processing*. Groningen: Grodil Press.

Hopp, H. (2009). The syntax-discourse interface in near-native L2 acquisition: off-line and on-line performance. *Biling. Lang. Cogn.* 12, 463–483. doi: 10.1017/s1366728909990253

Huang, C. T. J. (1982). Move wh in a language without wh movement. *Linguist. Rev.* 1, 369–416.

Johnson, A. M., Fiorentino, R., and Gabriele, A. (2016). Syntactic constraints and individual differences in native and non-native processing of Wh-movement. *Front. Psychol.* 7:549. doi: 10.3389/fpsyg.2016.00549

Johnson, J. S., and Newport, E. L. (1991). Critical period effects on universal properties of language: the status of subjacency in the acquisition of a second language. *Cognition* 39, 215–218.

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295x.99.1.122

Kim, E., Baek, S., and Tremblay, A. (2015). The role of island constraints in second language sentence processing. *Lang. Acquis.* 22, 384–416. doi: 10.1080/10489223.2015.1028630

Klein, E. C. (1993). *Toward Second Language Acquisition: A Study of Null-Prep*. Dordrecht: Kluwer Academic Publishers.

Kluender, R. (1991). *Cognitive Constraints on Variables in Syntax*. Ph.D. dissertation, University of California, San Diego, CA.

Kluender, R. (1998). "On the distinction between strong and weak islands: a processing perspective," in *The Limits of Syntax*, eds P. Culicover, and L. McNally (Leiden: Brill), 241–279.

Kluender, R., and Gieselman, S. (2013). "What's negative about negative islands? A re-evaluation of extraction from weak island contexts," in *Experimental Syntax and Island Effects*, eds J. Sprouse, and N. Hornstein (Cambridge: Cambridge University Press), 186–207. doi: 10.1017/cbo9781139035309.010

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

Kroch, A. (1981). On the role of resumptive pronouns in amnestying island constraint violations. *Chic. Linguist. Soc.* 17, 125–135.

Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *J. Psycholinguist. Res.* 25, 93–115. doi: 10.1007/bf01708421

Lewis, S., and Phillips, C. (2015). Aligning grammatical theories and language processing models. *J. Psycholinguist. Res.* 44, 27–46. doi: 10.1007/s10936-014-9329-z

Martohardjono, G. (1993). *Wh-Movement in the Acquisition of a Second Language: A Crosslinguistic Study of Three Languages with and without Overt Movement*. Doctoral dissertation, Cornell University, Ithaca, NY.

McCloskey, J. (1990). "Resumptive pronouns, A'-binding and levels of representation in Irish," in *The Syntax of the Modern Celtic Languages: Syntax and Semantics*, Vol. 23, ed. R. Hendrick (New York, NY: Academic Press), 199–248.

McCloskey, J. (2007). "Resumption," in *The Blackwell Companion to Syntax*, eds M. Everaert, and H. van Riemsdijk (Hoboken, NJ: John Wiley & Sons Ltd), 94–117.

McDaniel, D., and Cowart, W. (1999). Experimental evidence for a minimalist account of English resumptive pronouns. *Cognition* 70, B15–B24.

Meisel, J. M. (1997). The acquisition of the syntax of negation in French and German: contrasting first and second language development. *Second Lang. Res.* 13, 227–263. doi: 10.1191/026765897666180760

Morgan, A. M., and Wagers, M. W. (2018). English resumptive pronouns are more common where gaps are less acceptable. *Linguist. Inq.* 49, 861–876. doi: 10.1162/ling_a_00293

Nunes, J., and Uriagereka, J. (2000). Cyclicity and extraction domains. *Syntax* 3, 20–43. doi: 10.1111/1467-9612.00023

Omaki, A., and Schulz, B. (2011). Filler-gap dependencies and island constraints in second-language sentence processing. *Stud. Second Lang. Acquis.* 33, 563–588. doi: 10.1017/s0272263111000313

Pearl, L., and Sprouse, J. (2013). "Syntactic islands and learning biases: combining experimental syntax and computational modeling to investigate the language acquisition problem," in *Experimental Syntax and Island Effects*, Vol. 20, eds J. Sprouse, and N. Hornstein (Cambridge: Cambridge University Press), 23–68. doi: 10.1080/10489223.2012.738742

Perpiñán, S. (2010). *On L2 Grammar and Processing: The Case of Oblique Relative Clauses and the Null-Prep Phenomenon*. Ph.D. dissertation, University of Illinois at Urbana-Champaign, Champaign, IL.

Perpiñán, S. (2013). "Accounting for variability in L2 data: type of knowledge, task effects and linguistic structure," in *Innovative Research and Practices in Second Language Acquisition and Bilingualism*, ed. J. W. Schwieter (Amsterdam: John Benjamins Publishing Company), 165–192. doi: 10.1075/lllt.38.11per

Perpiñán, S. (2015). L2 grammar and L2 processing in the acquisition of Spanish prepositional relative clauses. *Biling. Lang. Cogn.* 18, 577–596. doi: 10.1017/s1366728914000583

Phillips, C. (2003). Linear order and constituency. *Linguist. Inq.* 34, 37–90. doi: 10.1371/journal.pone.0201700

Phillips, C. (2006). The real-time status of island phenomena. *Language* 82, 795–823. doi: 10.1016/j.cbpb.2018.09.006

Phillips, C. (2013). "On the nature of island constraints I: language processing and reductionist accounts," in *Experimental Syntax and Island Effects*, eds J. Sprouse, and N. Hornstein (Cambridge: Cambridge University Press), 64–108. doi: 10.1017/cbo9781139035309.005

Pickering, M. J., and Traxler, M. J. (1998). Plausibility and recovery from garden paths: an eye-tracking study. *J. Exp. Psychol. Learn.* 24, 940–961. doi: 10.1037/0278-7393.24.4.940

Pickering, M. J., Traxler, M. J., and Crocker, M. W. (2000). Ambiguity resolution in sentence processing: evidence against frequency-based accounts. *J. Mem. Lang.* 43, 447–475. doi: 10.1006/jmla.2000.2708

Polinsky, M., Clemens, E. L., Morgan, M. A., Xiang, M., and Heestand, D. (2013). "Resumption in English," in *Experimental Syntax and Island Effects*, eds J. Sprouse, and N. Hornstein (Cambridge: Cambridge University Press), 341–359. doi: 10.1017/cbo9781139035309.017

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Doctoral dissertation, MIT Press, Cambridge, MA.

Schachter, J. (1990). On the issue of completeness in second language acquisition. *Second Lang. Res.* 6, 93–124. doi: 10.1177/026765839000600201

Schwartz, B. D., and Sprouse, R. A. (1994). "Word order and nominative case in non-native language acquisition: a longitudinal study of (L1 Turkish) German interlanguage," in *Language Acquisition Studies in Generative Grammar*, eds T. Hoekstra, and B. D. Schwartz (Amsterdam: John Benjamins Publishing Company), 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

Schwartz, B. D., and Sprouse, R. A. (2000). "When syntactic theories evolve: consequences for L2 acquisition research," in *Second Language Acquisition and Linguistic Theory*, ed. J. Archibald (Oxford: Blackwell), 156–186.

Sells, P. (1984). *Syntax and Semantics of Resumptive Pronouns*. Doctoral dissertation, University of Massachusetts, Amherst, MA.

Shlonsky, U. (1992). Resumptive pronouns as a last resort. *Linguist. Inq.* 23, 443–468.

Slabakova, R. (2009). L2 fundamentals. *Stud. Second Lang. Acquis.* 31, 155–173. doi: 10.1017/s0272263109090263

Sprouse, J., and Hornstein, N. (2013). *Experimental Syntax and Island Effects*. Cambridge: Cambridge University Press.

Sprouse, J., Wagers, M., and Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language* 88, 82–123. doi: 10.1353/lan.2012.0004

Suñer, M. (1998). Resumptive restrictive relatives: a crosslinguistic perspective. *Language* 74, 335–364. doi: 10.1353/lan.1998.0194

Townsend, D. J., and Bever, T. G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Traxler, M. J., and Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: an eye-tracking study. *J. Mem. Lang.* 35, 454–475. doi: 10.1006/jmla.1996.0025

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

White, L. (2003). *Second Language Acquisition and Universal Grammar*. Cambridge: Cambridge University Press.

White, L., and Juffs, A. (1998). "Constraints on Wh-movement in two different contexts of nonnative language acquisition: competence and processing," in *The Generative Study of Second Language Acquisition*, eds S. Flynn, G. Martohardjono, and W. A. O'Neil (New York, NY: Lawrence Erlbaum Associates), 111–119.

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fpsyg.2020.00348

Intralingual Variation in Acceptability Judgments and Production: Three Case Studies in Russian Grammar

Anastasia Gerasimova[1,2]* *and Ekaterina Lyutikova*[1,2]

[1] Department of Theoretical and Applied Linguistics, Lomonosov Moscow State University, Moscow, Russia, [2] Pushkin State Russian Language Institute, Moscow, Russia

This paper contributes to the task of defining the relationship between the results of production and rating experiments in the context of language variation. We address the following research question: how may the grammatical options available to a single speaker be distributed in the two domains of production and perception? We argue that previous studies comparing acceptability judgments and frequencies of occurrence suffer from significant limitations. We approach the correspondence of production and perception data by adopting an experimental design different from those used in previous research: (i) instead of using a corpus we use production data obtained experimentally from respondents who are later asked to make judgments, (ii) instead of pairwise phenomena we examine language variation, (iii) judgments are collected formally using the conditions and materials from the production experiment, (iv) we analyze the behavior of each participant across the production and acceptability judgment experiments. In particular, we examine three phenomena of variation in Russian: case variation in nominalizations, gender mismatch, and case variation in paucal constructions. Our results show that there is substantial alignment between acceptability ratings and frequency of occurrence. However, the distribution of frequencies and acceptability scores do not always correlate. Speakers are not consistent in choosing a single variant across the two types of experiment. Importantly, the types of inconsistency they display differ, which means that the variation can be characterized from this point of view. We conclude that the degree of coherence of the two experiments reflects the effects of the evolution of variation over time. Another result is that elicited production and acceptability judgments vary with respect to how they reveal variation in language. In the case of the development or disappearance of variants, production indicates this earlier than judgments, and the rating task has the effect of restricting the choices available to respondents. However, the production method should not thereby be considered more sensitive. We argue that only a combination of production and judgment data makes it possible to estimate the directionality of changes in variability and to see the full distribution of different variants.

Keywords: acceptability judgments, gradience, production, experimental linguistics, variation, Russian

OPEN ACCESS

Edited by:
Susagna Tubau, Autonomous University of Barcelona, Spain

Reviewed by:
Alba Tuninetti, Bilkent University, Turkey
Maria Carmen Parafita Couto, Leiden University, Netherlands

***Correspondence:**
Anastasia Gerasimova
anastasiagerasimova432@gmail.com

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

Received: *30 July 2019*
Accepted: *14 February 2020*
Published: *31 March 2020*

Citation:
Gerasimova A and Lyutikova E (2020) Intralingual Variation in Acceptability Judgments and Production: Three Case Studies in Russian Grammar. Front. Psychol. 11:348. doi: 10.3389/fpsyg.2020.00348

INTRODUCTION

The idea that multiple sources of linguistic evidence provide complementary data is not novel. However, it still remains undetermined how different corpus and behavioral measures relate to each other. In this paper, we explore the correlation between the two linguistic domains of production and perception, by assessing the alignment between elicited production and acceptability judgments in the context of language variation.

Traditionally, acceptability judgments have served as the primary source of data for investigators engaged in developing linguistic theories. As the gathering of judgments has become more advanced (see Schütze, 1996; Featherston, 2007; Sprouse, 2007; among others) researchers have begun to use complex non-binary scales, such as the Likert scale. Consequently, the issue of the interpretation of gradience in judgment data has become more prominent. Although judgments are known to be gradient, it is not clear where this gradience comes from Phillips (2009), Schütze and Sprouse (2013), and Sprouse (2015). On one hand, gradience may result from factors other than grammar that affect language processing and decisions about acceptability, e.g., parser limitations and high working memory costs. Another option is that grammatical knowledge is itself gradient: combinations of different grammatical constraints lead to a range of grammaticality[1] levels.

Our assumptions about the grammatical architecture restrict our predictions with respect to different data sources. If grammar is considered categorical, gradience is reduced to an effect of extra-grammatical factors, i.e., processing mechanisms, which might differ in production and perception. Meanwhile, if grammar is gradient, we expect consistency in the data regardless of the source, be it judgments or produced texts. Consequently, the level of correspondence observed between the two language domains might shed light on what type of language modeling is preferable.

Our paper contributes to defining the relationship between production and perception by comparing the results of production and rating experiments in the context of language variation. We find two main problems with previous research on comparison between data sources. The first is that the production data used was retrieved from corpora. This approach has a serious drawback in that a particular selection of texts might not be comparable to the idiolects of the respondents giving their judgments. The second limitation is that the research was primarily focused on pairwise phenomena. This posits a conflict in terms of the dimensions of the data: while we expect a gradient scale of acceptability, we assume a binary choice in production. In this paper, we aim to provide a solution to both of these problems by analyzing the distribution of grammatical options in both the production and perception domains of individual speakers. In particular, we obtain both production and judgment data experimentally, using the same experimental conditions. Moreover, we examine three phenomena of variation in Russian, of the following type: variants are expected to exhibit different levels of acceptability, but none of them are prohibited in any particular context. Finally, we analyze the behavior of each participant individually, which helps us to understand the objective laws behind the data correspondence.

The rest of the paper is organized as follows. In section "Theoretical Background," we provide a brief overview of previous work on comparison between data sources, which includes the results of linking acceptability ratings with corpus data and other experimental methods. In section "The Present Study," we discuss the implicit assumptions behind the hypotheses tested in the previous research and formulate the objectives of the present study. This section also presents the materials for the experimental study – three types of constructions in Russian that display a certain degree of variability. In section "Experiments," we provide a description of the two series of experiments, involving production and judgments, conducted on the same sample of participants. In the same section, we estimate the level of correspondence between the two types of experiments by checking respondents' individual results. Section "Discussion of the Experimental Results" discusses the theoretical consequences of our findings. Final section concludes the study.

THEORETICAL BACKGROUND

Linking Acceptability Ratings and Corpus Data

Several recent studies investigate the relationship between acceptability judgments and frequency of occurrence. The main hypothesis is that grammatical knowledge is probabilistic and determines both frequency of occurrence and acceptability ratings. Consequently, on the basis of probabilities found in a corpus, one ought to be able to predict acceptability judgments. To formalize the gathering of these probabilities, investigators used language models that were fitted to the annotated corpus data in a supervised (Bresnan, 2007) or unsupervised manner (Lau et al., 2017; Sprouse et al., 2018).

Bresnan (2007) explored the correspondence between the two data sources with respect to the English dative alternation (e.g., *give the boy the book* vs. *give the book to the boy*). Using several contextual predictors, including various properties of the recipient and the theme, in the Switchboard corpus of spontaneous speech, the researcher created a statistical model that successfully predicted the choice of dative construction on the annotated test set. Then two experiments were conducted, which evaluated how the ratings provided by speakers correspond to the probabilities predicted by the model. The results showed that acceptability judgments corresponded to corpus probabilities. Even more importantly, linguistic manipulations with contextual predictors affected both probabilities and acceptability judgments in the same direction.

A conceptually different approach was proposed by Lau et al. (2014, 2015, 2017). In this study, acceptability judgments were predicted by unsupervised language models trained

[1] In line with Schütze (1996), we use the term *grammaticality* when referring to grammar as a mental construct, and the term *acceptability* when referring to judgments.

on raw text which did not contain any annotation or set predictors (in contrast with Bresnan, 2007). As likelihood of occurrence is partially determined by sentence length and lexical frequency, probabilistic language models were augmented with acceptability measures that compensate for additional frequency factors. The language models were tested on a dataset that contained sentences at varying levels of acceptability: original sentences retrieved from the British National Corpus and mappings of these sentences with errors introduced by round-trip machine translation[2]. Lau et al. (2017) then computed the Pearson correlation coefficient between the acceptability scores produced by computer models and mean human judgments. The comparison showed that some models achieved good levels of accuracy in predicting the observed gradient data. This result recommends these models as more effective than traditional formal grammars, which are unable to predict acceptability gradience at all.

A replication of this study was performed by Sprouse et al. (2018). The major criticism of the results from Lau et al. (2014, 2015, 2017) concerns the fact that round-trip translations might not create grammatical oppositions of the kind usually devised by syntacticians, whereby a specific grammatical property is manipulated while other properties remain constant in an experimental set. To formalize comparison between classical formal grammar and probabilistic language models with respect to accommodating gradient data, the datasets were enriched by randomly selected samples of pairwise and multi-condition phenomena. The results show that probabilistic models demonstrate a substantial loss in coverage of phenomena that are captured by categorical grammars and can be revealed in controlled syntactic experiments. In particular, the models fail to capture up to 35% of the phenomena that are accounted for in modern generative theory.

Notably, the three studies just mentioned relate acceptability judgments to production data retrieved from a corpus. This presupposes that the corpus embodies the grammatical constraints that are implied by speakers in rating tasks because all the text entries were produced by speakers of the same language. However, this approach has its limitations. Whether corpora correctly capture usage is still an open question. It is also essential to consider what types of texts are represented in corpora. While Bresnan (2007) used the Switchboard corpus of spontaneous speech, in the study by Lau et al. (2017) this factor was not controlled for, and Sprouse et al. (2018) used utterances from research literature. The problem is that data from texts belonging to particular genres might not be comparable to the results of acceptability tasks in which speakers are asked to evaluate the *naturalness* of the stimuli.

Another drawback concerns the type of data used in a language model: predictors identified by linguists, or features yielded in an unsupervised manner. Where a researcher uses predictors, it is doubtful whether all the predictors affecting the final result are in fact being distinguished. Additionally, it is unclear how to interpret findings at the lower end of the frequency spectrum. Testing on the basis of predictors is subject to limitations, as the corpus might lack all the possible combinations of predictor values that would be required by a comprehensive test. This problem was addressed by Divjak (2017), who analyzed *that*-clauses in Polish and encountered difficulties in determining which variables had an impact on acceptability ratings. Divjak (2017) suggests that implicit probabilistic syntactic knowledge is based not on n-gram frequency, but rather on higher-order knowledge (involving schemata or rules). However, the lack of any clear correspondence between frequency and ratings could result from the low capacity of the corpus.

The use of unsupervised language models is not trouble-free either. Language models take into account all kinds of information that can be retrieved from a corpus, which is not necessarily the same information that humans obtain when they acquire and use language. Thus, the replacement of the existing theoretical grammar models with computational ones would eliminate the explanatory function of language theory and modeling.

Taken together, the examined studies point toward the problem of corpus representativeness, which leads to flaws in the comparison between usage data and acceptability ratings. A possible solution would be to limit production data to the phenomenon under observation and obtain it specifically for the comparison at issue. In the following section, we review existing studies that have used a different source for production data, and provide the rationale for the present work.

Gathering Production Data Differently

A group of studies have approached the question of the correspondence between production and perception data by adding rigor to the production data gathering process. Instead of using language models trained on large datasets, the researchers obtained production frequencies in experiments.

To our knowledge, the first attempt to connect acceptability ratings to experimentally obtained production data was made by Adli (2011), who investigated the preferred subject position in Spanish *wh*-questions with respect to the thematic role of the *wh*-word. The database of elicited speech turned out to be rather limited: one expected option was completely absent. Hence, the representativeness of the database limited the potential for meaningful comparison.

The next study was carried out by Verhoeven and Temme (2017), who used a forced-choice task to evaluate results of production. They investigated the choice between SO and OS order in German clauses using two experimental procedures: forced-choice and split-100 rating. It was assumed that at some point in the production process, the speaker would compare a set of alternative expressions and judge their relative appropriateness in a particular context. This assumption is questionable as there is evidence that forced-choice is a form of rating task. Sprouse et al. (2018) reports that the results of forced-choice tasks, when transformed into ratings by means of the Elo system first developed for rating the relative strength of chess players, in fact correspond directly to the results of Likert scale tasks. The results

[2]The negative spectrum of acceptability was covered by sentences that were produced by first translating a set of sentences from the British National Corpus to four languages (Norwegian, Spanish, Japanese, and Chinese) and then translating them back to English using Google Translate.

of the experiments by Verhoeven and Temme (2017) turned out to be highly correlated. We think this is presumably due to the fact that speakers were ultimately carrying out the same rating task in both experiments[3].

Another attempt to relate acceptability ratings and elicited usage data was by Bermel et al. (2018), who retrieved probabilities from a balanced corpus and compared them to the distribution of existing options in fill-in-the-gap and rating tasks, completed by respondents simultaneously. Bermel et al. (2018) took the responses to the fill-in-the-gap task to serve as production data; however, they observed that this could more accurately be thought of as a forced-choice situation, as there were only two possible options in the two syntactic contexts. Although a correlation was found, there is a limitation to this study, namely, the performance of two distinct tasks within a single questionnaire. The main drawback of such a procedure is that the acceptability ratings could influence the production results and vice versa.

To summarize this brief review, we argue that previous comparisons of acceptability judgments and production based on the information retrieved from corpora have the following limitations. First, the corpus data may incorrectly represent the speech of the respondents providing judgments, due to differences in the text types involved. Second, speech corpora give rise to difficulties in dealing with low frequency spectrum phenomena. Third, the use of probabilistic language models raises the issue of model parameters. Where predictors have been pre-defined by linguists it is unclear whether the whole range of predictors affecting the final result has been taken into consideration. In the case of unsupervised feature detection, the algorithm may use all kinds of information that can be retrieved from the corpus, which is not necessarily the information that humans obtain when they acquire and use language. Finally, those studies which aimed to control for relevant factors when gathering production data did not change the overall picture. Intrinsically, these studies were comparing different acceptability rating methods and considering how well their results correspond to the predictions of probabilistic language models. In the next section we suggest how the research question can be modified to overcome these limitations.

THE PRESENT STUDY

Before we formulate the objectives of the present study, we would like to discuss the implicit assumptions behind the hypotheses tested in the previous research. The fundamental idea concerns the nature of grammatical knowledge: if grammar is probabilistic, it determines both offline production and comprehension, which are externalized quantitatively in frequency of occurrence and acceptability ratings, respectively. However, it is essential to ascertain what kind of grammatical knowledge is presupposed in this approach. In most of the studies discussed above, production data was retrieved from a corpus and was compared to acceptability ratings provided by a group of speakers. That is, the relevant instances of production were determined by the grammatical knowledge of the individuals who produced the set of texts that happened to be included in the corpus. Production data in this case reflects 'collective grammar,' which is not necessarily reducible to a simple sum of individual idiolects (Bailey, 1973; Bickerton, 1975; Wolfram and Beckett, 2000; Kuhl, 2003) and represents the individual grammars only to the extent of what is present in the texts. Meanwhile, judgment data is determined by the grammatical knowledge of speakers who participate in the survey, representing another form of 'collective grammar.' The question therefore arises as to whether investigators are comparing entities of the same nature when looking for correlations between frequencies and ratings.

In general, it is presupposed that an individual belonging to a language community possesses the same grammar as the people with whom she communicates – that is, other members of the same community or social group (Horvath and Sankoff, 1987). This methodology is based on the "homogeneity assumption" that individual-speaker variation is not important in describing variation in general (Wolfram and Beckett, 2000). In the reviewed research a conceptually similar idea is assumed, namely that those speakers who participate in the surveys possess the same grammar as those who composed the texts found in the corpus. However, this assumption is untenable because it is possible that the language community providing the production frequencies and the individuals providing the ratings possess grammars that are far from equal. In other words, using a corpus means that an additional factor needs to be taken into account: the level of coherence between the grammar of the survey participants and the collective grammar reflected in the corpus.

We suppose that if grammatical knowledge is indeed probabilistic, one would see consistent patterns in the production and comprehension of a single speaker, without the mediation of the collective language system of all speakers or speakers from a certain community. Our prediction is that in this case there would be a one-to-one correspondence between the production data and acceptability judgments of a given speaker. Both production and judgment data should provide the same ranking of variants: the most frequent variant would also be the most acceptable, and the least frequent variant would be the least acceptable.

Another important issue is connected to the type of phenomena on the basis of which the two language domains were compared. In most of the studies mentioned, linguists analyzed alternations that were dependent on a set of contextual predictors, distinguished and annotated by investigators in advance. This means that there were contexts where one alternative was acceptable while the other was not. Although the question regarding the completeness of the set of predictors

[3]A similar comparison of methods was reported in Klavan and Veismann (2017), which compared the performance of a corpus-based language model against the results of two rating experiments: forced-choice and Likert scale. The results of the study show that forced-choice data provides a slightly better reflection of the corpus than Likert scale data. This might result from the higher statistical power of the method. For instance, Stadthagen-González et al. (2017) compare Likert scale judgments to a 2-alternative forced-choice task combined with the Thurstone measurement model, which allows the results of comparisons to be laid out along a single interval scale. The results of the study show that with some experimental conditions a forced-choice task might yield more granular data than pure ratings.

remains open, the distribution of predictors might dictate the quantitative values for frequency of variants. Consequently, the ratings for a certain phenomenon are not directly compared to the distribution of that phenomenon but instead to the distribution of predictors that favor a certain value. The relation between the distribution of predictors and the distribution of variants might be non-linear, at least because not all the predictors are distinguished and there may be interaction between them. We do not aim to explore the nature of this relationship; our point is that such an approach lends additional complexity to any hypothesis about the relation between offline perception and production.

The analysis of pairwise phenomena, as in Lau et al. (2017) and Sprouse et al. (2018), is also insufficient. It presupposes a binary distribution of language data: (i) without any violations of functional/grammatical constraints, (ii) with such violations. In this case, the comparison is carried out between variables of different dimensions: in production there is a binary choice, between producing and not producing a construction, while in perception there is a scale of acceptability.

To avoid the issue of predictors and problems arising from the binary distribution of language data, we suggest studying phenomena that supposedly exhibit free variation: although variants may favor certain contexts, none of them seem to violate any constraint and thus to be unacceptable in any particular context.

Therefore, in the present study, we use a hypothesis on the correspondence between offline production and comprehension that requires fewer assumptions than the hypotheses used in previous research. We address the following research question: how are the grammatical options distributed in both the production and perception domains of a single speaker? We believe that answering this question will contribute to the task of connecting gradient acceptability judgments and usage, as it eliminates the problems of corpus representativeness and binary opposition in the language phenomena under examination.

We approach the correspondence of production and perception data by adopting an experimental design alternative to those used in previous research. Firstly, instead of using a corpus we use production data obtained experimentally from respondents who are later asked to make judgments. Secondly, instead of pairwise phenomena we examine language variation. The phenomena that we explore include those involving more than two alternatives, to the effect that we do not end up with a forced-choice task when gathering production data. Thirdly, judgments are collected formally using the conditions and materials from the production experiment. Finally, we analyze the behavior of each participant across the production and acceptability judgment experiments.

The Phenomena Under Observation

We examine three phenomena of variation in Russian. The choice of phenomena was premised on the status of variation: we aimed to use both data with predictors and data with free variation. This way we could replicate the choice of data from both types of study undertaken previously: those that used data with annotated predictors, and those that used raw data[4].

The first phenomenon addressed is **case variation in nominalizations.** Russian event nominalizations belong to the ergative-possessive type (Koptjevskaja-Tamm, 1993), which means that arguments of intransitives and internal arguments of transitive stems are marked with the possessive, genitive case (GEN), while external arguments of transitives are assigned instrumental case (INSTR). However, for some stems the external argument can be marked both GEN and INSTR: this is possible for nominalizations with a lexically governed internal argument (1) and for nominalizations derived from unergative stems (2). That is, the case marking strategy is one of the parameters of intralingual variation for Russian.

(1) a. *torgovlja* *fermera* *skotom*
trading farmer.GEN cattle.INSTR
b. *torgovlja* *fermerom* *skotom*
trading farmer.INSTR cattle.INSTR
'trading in cattle by the farmer'

(2) [Gerasimova, 2016: (8)]
a. *Gracioznoe* *xoždenie* *modelej* . . .
graceful walking of the models.GEN
b. *Gracioznoe* *xoždenie* *modeljami* ***po podiumu*** . . .
graceful walking by the models.INSTR **on the runway**
. . . *bylo vysoko oceneno dizajnerom.*
'Graceful walking of the models (on the runway) was highly appreciated by the designer.'

The case marking strategy depends on the structural properties of the nominalization: thus, adverbial PP modification increases the acceptability of INSTR (2) (Pereltsvaig, 2017), an observation supported by the experimental data from Pereltsvaig et al. (2018). This aspect is modeled within the framework of formal syntax in terms of the amount of structure that is nominalized: the syntactic structure is claimed to be more complex when an adverbial PP is merged, which makes it similar to the structure of transitives. Pereltsvaig (2017) connects the larger structure of nominalization with the licensing of INSTR. This means that even when there is no adverbial PP modification of a nominalization with a lexically governed internal argument, but its external argument is nonetheless marked INSTR, the nominalization is supposed to possess a larger structure. If we rely on the theoretical modeling proposed by Pereltsvaig, we might suppose that in the absence of a PP the smaller structure would be preferred on the basis of Economy Principle considerations. Therefore, a general preference for GEN is expected for both production and acceptability judgments. With respect to our

[4]Another advantage in using the examined set of phenomena is that none of them are mentioned in prescriptive grammars. This means that respondents were not influenced by prescriptive grammars and would not draw on their school knowledge of grammar when participating in the experiments.

goals, this phenomenon presents variation with binary choice and no identified predictors.

The second phenomenon is **gender mismatch**, which occurs in the context of masculine nouns that denote a professional status and refer to females. These nouns can trigger both masculine and feminine agreement on attributive modifiers and past tense verbs (Muchnik, 1971; Crockett, 1976; Shvedova, 1980; Pesetsky, 2013; Lyutikova, 2015; among others). The three possible agreement patterns are: GRAMMATICAL AGREEMENT, where all agreeing constituents are masculine (3a), REFERENTIAL AGREEMENT, within which modifiers are masculine and the verb is feminine (3b), and REFERENTIAL ATTRIBUTIVE AGREEMENT, where non-classifying adjectives [adjectives without an idiomatic interpretation (Rothstein, 1980; Svenonius, 2008; Pesetsky, 2013)] and the verb are feminine (3c). The majority of investigators suggest that the observed variation results from a process of "feminization" at some stage in the derivation, which henceforth determines the agreement pattern of the nominal (Pereltsvaig, 2006, 2015; Asarina, 2009; Pesetsky, 2013; Lyutikova, 2015; Puškar, 2017; Steriopolo, 2018; and others). To date, no specific factors have been identified as influencing the choice of agreement pattern. REFERENTIAL AGREEMENT is assumed to be the most frequent pattern in actual usage: consequently, we would expect it to be the most used and the most acceptable pattern in the experiments. This variation presents multiple agreement choices, not limited to binary distribution: the three mentioned patterns are all considered acceptable by both traditional grammars and formal syntactic studies.

(3) a. GRAMMATICAL AGREEMENT pattern: all agreeing constituents are masculine.

nov-yj	*zubn-oj*	*vrač*	*prišel*
new-M	dental-M	doctor.M	arrived-M

b. REFERENTIAL AGREEMENT: modifiers are masculine, the verb is feminine

nov-yj	*zubn-oj*	*vrač*	*prišl-a*
new-M	dental-M	doctor.M	arrived-F

c. REFERENTIAL ATTRIBUTIVE AGREEMENT: non-classifying adjectives and the verb are feminine.

nov-aja	*zubn-oj*	*vrač*	*prišl-a*
new-F	dental-M	doctor.M	arrived-F

d. ILL-FORMED pattern: non-classifying adjective is feminine but the verb is masculine.

* *nov- aja*	*zubn-oj*	*vrač*	*prišel*
new-F	dental-M	doctor.M	arrived-M

'the new dentist arrived'

The third phenomenon is **case mismatch in paucal constructions.** In paucal constructions feminine nominalized adjectives and adjectives that modify feminine nouns can be marked either NOM or GEN (4)–(5) (Graudina et al., 1976; Shvedova, 1980; Golub, 1997; and others). The choice of case marking partially depends on the context of the paucal construction: NOM is preferred in argumental (DP) position, where the paucal construction agrees with the predicate, and GEN is used primarily in quantificational (QP and PP) positions, where there is no predicate agreement (Shkapa, 2011; Lyutikova, 2015). Corpus studies by Shkapa (2011) and Gerasimova (2019) have shown that in general the NOM form is more frequent in paucal constructions. Therefore, NOM is expected to be the preferred option in both production and perception experiments. Some previous studies also claim that the choice of case marking on the adjectival constituent depends on internal properties of the paucal construction, such as the morphological type of the adjective or stress position on the noun. However, according to Shkapa (2011) there is no evidence for these predictions. We consider this variation to have an identified predictor, namely, the presence of predicate agreement.

(4) a. *dve* *gornič-yje /* *gornič-yx*
two maid(F)-NOM.PL / maid(F)-GEN.PL
'two maids'

b. *tri* *dobr-yje /* *dobr-yx* *devuški*
three kind-NOM.PL / kind-GEN.PL girls.F
'three kind girls'

(5) a. **DP context. Agreement with predicate.**

[*Dve* *gornič-yje /* *gornič-yx*]
two maid(F)-NOM.PL / maid(F)-GEN.PL
ubirali nomer *k priezdu gostja.*
tidied the room before guest's arrival.
'Two maids tidied the room before the guest's arrival.'

b. **PP context. Comparative construction.**

Etot vypusk ***na*** [*tri*
This issue is **PREP** three
jark-ije / *jark-ix* *kartinki*]
bright-NOM.PL / bright-GEN.PL pictures.F
bogače, chem včerašnij.
richer than yesterday's.
'This issue is three bright pictures richer than yesterday's.'

c. **PP context. Distributive construction.**

Každaja *vypusknica* *možet* *priglasit'* ***po***
each graduate can invite **PREP**
[*dve* *znakom-yje /* *znakom-yx*]
two friend(F)-NOM.PL / friend(F)-GEN.PL
'Each graduate can invite at most two acquaintances.'

d. **QP context. Impersonal predicate, no agreement.**

Na stole ***ostalos'*** [*tri* *igral'n-yje /*
On the table **left.IMPRS.PST** three playing-NOM.PL /
igral'n-yx *karty*]
playing-GEN.PL cards.F
'On the table there remained three playing cards.'

To sum up, we have chosen three phenomena that differ with respect to the type of variation they display. Firstly, in all three cases two or more variants are acceptable and none of the variants explicitly violates any functional or grammatical constraints. Nonetheless, there are some predictions with respect

to the most frequent option (case or agreement pattern). Secondly, the variation is not fully determined by predictors. Only in the case of paucal constructions are contextual predictors identified, although their presence does not guarantee any particular choice (as shown in Shkapa, 2011). In the case of nominalizations, variation can be manipulated by adding an adverbial PP into the structure; when there is no PP the variation is considered to be free. It is not known how gender mismatch can be manipulated either. Finally, it may be that the variants are distributed unequally over speakers. In particular, in Pereltsvaig et al. (2018) it was shown that some speakers are consistent in using both GEN and INSTR, while some do not allow INSTR at all. There is no similar data for gender mismatch and paucal constructions; however, it is possible that these two phenomena are also characterized by a cross-speaker distribution of variants. This property of the variation should not influence the hypothesis testing, as in case there is any intraspeaker variation we would expect a speaker to be consistent in her choices in both perception and production.

The reviewer raised the issue of the linguistic comparability of the phenomena with respect to their source. The three phenomena under discussion appear to be grammatically comparable due to the uniformity of the syntactic structures and mechanisms behind feature interpretation and valuation within a given language (Adger and Svenonius, 2011)[5]. All three involve variation that arises in the process of feature valuation with respect to the constituent that enters derivation bearing an unvalued feature. Variation results from the fact that there is more than one controller available for feature valuation: the gender agreement controller in case of gender mismatch, and the case governor for nominalizations and paucal constructions. The availability of multiple controllers may be inherent to the structure (as in paucal constructions) or originate from conscious or subconscious structure varying (as in the case of gender mismatch and nominalizations, respectively). On the basis of these observations we suppose that the three investigated phenomena can be attributed to the same component of grammar, namely, narrow syntax, and can be assumed to involve the same type of grammatical operation, viz., feature valuation.

EXPERIMENTS

In order to investigate the correspondence between the distribution of grammatical options in both offline production and offline perception of a single speaker we conducted a series of linguistic experiments using the three Russian phenomena presented above. For each phenomenon, we carried out two experiments: a production experiment, in which respondents were asked to provide the case/agreement morphology themselves, and an acceptability judgment experiment, in which respondents provided acceptability judgments using a 5-point Likert scale. We first conducted the three production experiments, one for each phenomenon; then 5 months later the three judgment experiments were launched. In both sets of experiments, we made use of the same group of participants. We suppose that the chosen period between the sets of experiments was long enough to eliminate any syntactic satiation effect. In addition, as we were using the same materials in both it was necessary that the speakers forget the stimuli in the intervening period. We assume that a span of several months is sufficient to achieve both goals: however, there is more to be done with respect to defining the proper timing for such a series of experiments[6]. When participating in a set of experiments, respondents completed separate experiments in one day with breaks half an hour long in between: the respondent first completed the nominalization experiment, then the gender mismatch experiment, and finally the experiment on paucal constructions. All participants encountered the experiments in the same order. The breaks between experiments were arranged in order to avoid fatigue effects.

Participants

One hundred and ten self-reported native Russian speakers participated in the three production experiments (82 females). Ages ranged from 15 to 49 (mean age 21, SD 5.3). Fifty-eight of these participants subsequently completed the three acceptability judgment surveys (43 females). This time ages ranged from 17 to 37 (mean age 21, SD 4.7). All participants provided informed consent and were naïve as to the purpose of the study and the research question. The experiments were carried out in accordance with the Declaration of Helsinki and the existing international regulations concerning ethics in research. The participants performed the task remotely, via the web-based software Google Forms. Participants were presented with one sentence at a time; the time allowed for the answer was not limited but participants were instructed to complete the task as fast as possible.

Materials and Procedure

In this section, we discuss experimental materials for each phenomenon. We first review the experimental factors and the number of stimuli in both production and acceptability judgment experiments. Then, we describe the sample stimuli and the production task. In all production experiments, the task for respondents was to provide the case or agreement morphology, and the only differences concern how the material to be filled in was presented. The section ends with a discussion of the item-to-filler ratio, the training sentences and the procedure involved in the acceptability judgment experiments. In all the experiments

[5]We thank the reviewer for bringing this issue to our attention.

[6]To the best of our knowledge there are no studies exploring how long it takes for respondents to forget linguistic stimuli. The effects of repeated exposure of linguistic stimuli have only been studied with respect to syntactic satiation, an effect whereby sentences that were initially judged ungrammatical come to be judged as acceptable. However, this phenomenon is usually studied within a single testing session (as in Francom, 2009; Hofmeister et al., 2013).

reported in this study counterbalancing was achieved by means of pseudorandomization and a Latin square design.

In the case of **nominalizations**, there was only one factor in the production experiment – the type of nominalized verbal stem. These are transitive stems with lexically governed internal argument and unergatives, for which we expected variation, versus 'normal' transitives and unaccusatives, for which we expected no variation and that were used as baseline conditions. We constructed 16 target sentences, four for each of the four conditions. The target sentences were presented in four pseudorandomized orders and interspersed with 32 filler items of comparable structure and length, which contained participles instead of nominalizations.

In each condition from the production experiment, there was a choice between GEN and INSTR. Therefore, in the acceptability judgment experiment one more factor was added, namely, the case marking of the external argument. The number of stimuli from the production experiment was multiplied by two, giving 32 sets of target sentences in the judgment experiment. We used the 16 sets of stimuli that had already been used and added 16 more sets (see **Supplementary Data Sheet 1** for production experiment stimuli). Sample stimuli from the **Table 1** represent one set. The number of filler sentences was kept the same in order to avoid fatigue effects.

Each stimulus was constructed in the following manner: the first part of the sentence contained the finite verb with its arguments, and the second part contained the nominalization formed from that verb. In the production experiment, speakers were asked to generate arguments of nominalizations, assigning the case that sounded most natural to them in each instance. The second conjunct of a complex sentence contained a gap which the participant had to fill in with the argument from the preceding context (the first conjunct of the sentence) (6).

(6) V tot mesjac **armija** osvobodila
That month the army.NOM reconquered

stolicu, i osvoboždenie _____ _____
the capital.ACC, and reconquest _____ _____

sil'no podnjalo boevoj dux soldat.
greatly lifted the martial spirit of the soldiers.

(To fill in: of the capital by the army).

In the **gender mismatch** experiment, we examined gender agreement for various combinations of adnominals (determiners: possessive and demonstrative pronouns; high adjectives; low adjectives). All eight combinations that were used are listed in (7).

(7) a. det high adj. low adj. our hard-working executive **supervisor** organized
b. det high adj. our hard-working **supervisor** organized
c. det low adj. our executive **supervisor** organized
d. det our **supervisor** organized
e. high adj. low adj. hard-working executive **supervisor** organized
f. high adj. hard-working **supervisor** organized
g. low adj. executive **supervisor** organized
h. (no adnominals) **supervisor** organized

det = determiner (possessive/demonstrative pronoun).

Each combination from (7) was used twice in the experiment, which yields 16 sets of experimental sentences (see the sample stimuli in **Table 2**). Thirty-two filler items contained nouns that unambiguously denote the sex of the referent.

TABLE 1 | Conditions from the nominalization experiments.

Condition	Type of nominalized stem	Case of external argument (judgment experiment only)	Example
1–2	Transitive	GEN-INSTR	V tot mesjac **armija** osvobodila **stolitcu**, i osvoboždenie **armii/armiej** stolicy sil'no podnjalo boevoj dux soldat. That month **army.NOM** reconquered **capital.ACC**, and reconquest **army.GEN/army.INSTR capital.GEN** greatly lifted the martial spirit of the soldiers.
3–4	Transitive with lexically governed internal argument	GEN-INSTR	V techenie matča **sud'ja** podygryval **komande**, a podygryvanie **sud'i/sud'ej komande** strogo zapreščeno po pravilam čempionata. During the game **referee.NOM** favored **team.DAT,** and favoring **referee.GEN/referee.INSTR team.DAT** is strictly prohibited by the championship rules.
5–6	Unergative	GEN-INSTR	Posle procedury **pacient** stal kašljat', i kašljanie **pacienta**/**pacientom** srazu nastorožilo lečaščego vrača. After the procedure **patient.NOM** began to cough, and coughing **patient.GEN/patient.INSTR** immediately attracted the doctor's attention.
7–8	Unaccusative	GEN-INSTR	Každuju osen' **babuška** priezžala k nam v gorod, i priezd **babuški/babuškoi** vsegda soprovoždalsja vkusnym i sytnym zastol'em. Every autumn **grandmother.NOM** arrived in the city, and arrival **grandmother.GEN/grandmother.INSTR** was always followed by a holiday feast.

TABLE 2 | Conditions from the gender mismatch experiments.

Condition	Adnominals in NP	Agreement pattern (judgment experiment only)	Example
1	Det High Low	GRAMMATICAL	Vsju noch' Tane ne udalos' somknut' glaz: **nash otvetstvennyj proektnyj** menedzher **gotovil** prezentaciju reklamnoj kampanii dlja radioholdinga. Tanja couldn't get a wink of sleep all night: **our.M responsible.M project.M** manager **was preparing.M** the presentation of a promotional campaign for the radio corporation.
2		REFERENTIAL	**our.M responsible.M project.M** manager **was preparing.F**
3a		REFERENTIAL ATTRIBUTIVE	**our.F responsible.M project.M** manager **was preparing.F**
3b		REFERENTIAL ATTRIBUTIVE	**our.F responsible.F project.M** manager **was preparing.F**
4		ILL-FORMED	**our.M responsible.F project.M** manager **was preparing.M**
5	Det High	GRAMMATICAL	**our.M responsible.M** manager **was preparing.M**
6		REFERENTIAL	**our.M responsible.M** manager **was preparing.F**
7a		REFERENTIAL ATTRIBUTIVE	**our.F responsible.M** manager **was preparing.F**
7b		REFERENTIAL ATTRIBUTIVE	**our.F responsible.F** manager **was preparing.F**
8		ILL-FORMED	**our.M responsible.F** manager **was preparing.M**
9	Det Low	GRAMMATICAL	**our.M project.M** manager **was preparing.M**
10		REFERENTIAL	**our.M project.M** manager **was preparing.F**
11		REFERENTIAL ATTRIBUTIVE	**our.F project.M** manager **was preparing.F**
12		ILL-FORMED	**our.M project.F** manager **was preparing.M**
13	Det	GRAMMATICAL	**our.M** manager **was preparing.M**
14		REFERENTIAL	**our.M** manager **was preparing.F**
15		REFERENTIAL ATTRIBUTIVE	**our.F** manager **was preparing.F**
16		ILL-FORMED	**our.F** manager **was preparing.M**
17	High Low	GRAMMATICAL	**responsible.M project.M** manager **was preparing.M**
18		REFERENTIAL	**responsible.M project.M** manager **was preparing.F**
19		REFERENTIAL ATTRIBUTIVE	**responsible.F project.M** manager **was preparing.F**
20		ILL-FORMED	**responsible.M project.F** manager **was preparing.M**
21	High	GRAMMATICAL	**responsible.M** manager **was preparing.M**
22		REFERENTIAL	**responsible.M** manager **was preparing.F**
23		REFERENTIAL ATTRIBUTIVE	**responsible.F** manager **was preparing.F**
24		ILL-FORMED	**responsible.F** manager **was preparing.M**
25	Low	GRAMMATICAL	**project.M** manager **was preparing.M**
26		REFERENTIAL	**project.M** manager **was preparing.F**
27		REFERENTIAL ATTRIBUTIVE	**project.F** manager **was preparing.F**
28		ILL-FORMED	**project.F** manager **was preparing.M**
29	No	GRAMMATICAL	manager **was preparing.M**
30		REFERENTIAL	manager **was preparing.F**

In the judgment experiment, four patterns were examined for each combination: GRAMMATICAL AGREEMENT, REFERENTIAL ATTRIBUTIVE AGREEMENT, REFERENTIAL AGREEMENT, and ILL-FORMED AGREEMENT patterns. Two important properties of the stimuli must be pointed out. Firstly, for combination (7h) only two agreement patterns were logically available (GRAMMATICAL AGREEMENT and REFERENTIAL AGREEMENT). As shown in **Table 2**, conditions 29 and 30 correspond to this combination. Secondly, for combinations (7a) and (7b) the REFERENTIAL ATTRIBUTIVE AGREEMENT pattern could be applied in two ways: either only the determiner demonstrates feminine agreement and the high adjective remains masculine, or both determiner and high adjective are feminine. Pesetsky (2013) considers both variants to be equally acceptable; in contrast, Pereltsvaig (2015) predicts that the two adnominals cannot be mismatched. As there is no agreement between investigators and no experimental data that would provide evidence for either point of view, we introduced the two possibilities as two separate conditions: conditions 3a and 3b, 7a and 7b in **Table 2** for combinations (7a) and (7b), respectively. Consequently, the two factors, combination and agreement pattern, adjusted according to the considerations mentioned above give 32 conditions in total. In each experiment, there were two sentences for each condition. Fillers were the same as in the production experiment. We chose these quantities of target and filler items in order to avoid fatigue effects.

The target items were complex sentences, in which the first clause provided a context that explicitly indicated the gender of the human denoted by the subject in the second coordinate clause. This was done by using traditionally female names. This part of the sentence involved no agreement morphology. The second clause contained a noun phrase and a verb in the past tense, with gaps instead of endings in the production experiment. Speakers were asked to write the attributive modifiers and

the verb with the endings in the textbox so that the sentence was complete (8).

(8) Vsju noč Tane ne udalos' somknut' glaz:
nash_ otvetstvenn_ proektn_ menedžer **gotovil_**
prezentaciju reklamnoj kampanii dlja radioholdinga.
Tanja (female name) couldn't get a wink of sleep all night: ***our responsible project*** *manager* ***was preparing*** *the presentation of a promotional campaign for the radio corporation.*

(9) a. nash otvetstvennyj proektnyj gotovil
our-M responsible-M project-M was preparing-M
b. nash otvetstvennyj proektnyj gotovila
our-M responsible-M project-M was preparing-F

In the **paucal constructions** production experiment, we controlled for context (QP, DP, and PP), animacy, and pattern, i.e., whether the paucal construction involved feminine nominalized adjectives or modified feminine nouns. This gives 12 conditions in total. With two sentences for each condition there were 24 sets of target sentences. The sentences were kept relatively short. The target sentences were interspersed with 48 filler items of comparable structure and length, which contained numeral constructions involving other numerals and nouns of different grammatical genders.

The acceptability judgment experiment involved one more factor – case: in each condition from the production experiment there was a choice between NOM and GEN. Therefore, the number of stimuli in the judgment experiment was multiplied by two in comparison to the production experiment (see **Table 3** for the sample set of stimuli). Filler items were kept the same.

In the production experiment, the task was to inflect a paucal construction whose component parts (numeral + noun phrase) were provided in parentheses. The numeral was represented with a digit from 2 to 4, and alongside it there was either a nominalized adjective [as in example (10)], or a noun modified by an adjective, given in the singular. The rationale behind this choice is that in Russian paucal constructions the form of the modifying adjective is plural. The form was given in the singular because otherwise we would have to give the NOM.PL, which might lead respondents to prefer that over the GEN.PL and cause a priming effect.

(10)_____ (2, pračečnaja) byli otremontirovany
_____ (2, laundry(F)-NOM.SG) have been renovated
v etom mesjace.
this month.

All production tasks were designed so that participants could give only one answer. Only one phenomenon out of three presupposed a binary distribution of answers (namely, nominalizations, where respondents had to choose GEN or INSTR). In the gender mismatch experiment respondents could

TABLE 3 | Conditions from the paucal construction experiments.

Condition	Context	Pattern	Animacy	Case	Example
1–2	DP	Nominalized adjective	Animate	NOM-GEN	Dve **beremennye/beremennyx** obsuždali novosti sidja na skamejke. Two **pregnant woman(F)-NOM.PL/pregnant woman(F)-GEN.PL** were discussing the news sitting on a bench.
3–4	DP	Nominalized adjective	Inanimate	NOM-GEN	Dve **pračečnyje/pračečnyx** byli otremontirovany v gorode v etom mesjace. Two **laundry.NOM.PL/laundry.GEN.PL** have been renovated in the town this month.
5–6	DP	Noun	Animate	NOM-GEN	Tri **veselye/veselyx devočki** obsuždali plany na vyxodnye. Three **cheerful-NOM.PL/cheerful-GEN.PL girls** were discussing plans for the weekend.
7–8	DP	Noun	Inanimate	NOM-GEN	Dve **sočnye/sčnyx gruši** byli ostavleny v novoj vaze. Two **juicy-NOM.PL/juicy-GEN.PL pears** were left in a new bowl.
9–10	QP	Nominalized adjective	Animate	NOM-GEN	Včera za etot srok prinjato dve **beremennye/beremennyx**. Yesterday in the same period an appointment was given to two **pregnant woman(F)-NOM.PL/pregnant woman (F)-GEN.PL**.
11–12	QP	Nominalized adjective	Inanimate	NOM-GEN	V etom rajone za god obustroeno **dve pračečnye/pračechnyx.** In this neighborhood within a year there were equipped two **laundry.NOM.PL/laundry.GEN.PL.**
13–14	QP	Noun	Animate	NOM-GEN	V sledujuščii etap viktoriny prošlo dve **veselye/veselyx devočki.** Into the next stage of the quiz were accepted three **cheerful-NOM.PL/cheerful-GEN.PL girls.**
15–16	QP	Noun	Inanimate	NOM-GEN	Na stole k večeru ostalos' dve **sočnye/sočnyx gruši.** On the table by the end of the day there remained two **juicy-NOM.PL/juicy-GEN.PL pears**.
17–18	PP	Nominalized adjective	Animate	NOM-GEN	Za každyi čas vrač prinimaet po dve **beremennye/beremennyx.** Every hour the doctor gives an appointment to two **pregnant woman(F)-NOM.PL/pregnant woman (F)-GEN.PL**.
19–20	PP	Nominalized adjective	Inanimate	NOM-GEN	V každom rajone kompanija otkryla po dve **pračečnye/pračečnyx.** In every neighborhood the company opened two **laundry.NOM.PL/laundry.GEN.PL.**
21–22	PP	Noun	Animate	NOM-GEN	Na každuju lavočku režisser posadil po tri **veselye/veselyx devočki.** On every bench the director seated three **cheerful-NOM.PL/cheerful-GEN.PL girls.**
23–24	PP	Noun	Inanimate	NOM-GEN	Každomu gost'u xozjajka vydala po dve **sočnye/sočnyx gruši.** To every guest the hostess gave two **juicy-NOM.PL/juicy-GEN.PL pears**.

choose from multiple variants, all of which were restricted to the phenomenon in question, and in the experiment on paucal constructions respondents could choose alternative constructions (the interpretation of digits was not restricted, so respondents could use collective numerals or quantificational nouns; the latter were chosen in 5.33% of responses). The risk we were running with the nominalization experiments was that we would end up with a forced-choice task. However, as was discussed above (see section "Gathering Production Data Differently"), forced-choice should be considered a rating task: therefore, we would not expect any differences in the results between the production and acceptability judgment experiments.

The procedure for all the acceptability judgment experiments was the same. Respondents were asked to rate each sentence on a scale from 1 to 5, where 1 represents *bad* or *unnatural* and 5 represents *good* or *natural*[7]. Participants were told that the task had no correct answers and had nothing to do with what is advocated in prescriptive grammar or the plausibility of the described event.

The first four trials in each experiment served as training sentences and were identical for all participants. Out of the 110 respondents who completed the survey, four participants were excluded as they did not understand the task, yielding 106 participants whose data was later analyzed. As in the production experiments, at the beginning of the judgment experiments there were four training sentences, which provided grounds for excluding any participants who did not provide judgments at the expected end of the spectrum[8]. On the basis of this metric we excluded 1 participant out of the 58 who completed the surveys, which yields 57 participants whose data was later analyzed.

The described quantitative properties of the stimuli from the experiments are presented in **Table 4**. These numbers and, consequently, the number of stimuli responded to remain the same for all the participants despite the individual results in the production experiments. The number of filler items was adjusted to eliminate fatigue effects: when the number of target sentences was less than 25, the item-to-filler ratio was 1:2, and when there were more the item-to-filler ratio was 1:1. The general principle was not to exceed a total of 100 sentences, giving a survey that could be completed in approximately 15–20 min.

It is important to note that not all the controlled variables were independent variables, i.e., involved in the hypothesis testing. In particular, the combinations of adnominals from the gender mismatch experiment and animacy in the paucal construction experiment were considered extraneous variables, i.e., they were not intentionally tested in the experiments, but they were controlled for, as there was a possibility that they could influence the final results[9]. There was no effect found for these two variables[10].

Data Analysis

All production experiments in the paper were analyzed by means of the same data analysis procedure: each experiment involved from one to three explanatory variables or predictors, and we observed a categorical response with 2 or more values. Therefore, the data from the production experiments were fitted to a logistic regression model (Levshina, 2015) with the following factors: the stem type in the case of nominalizations; number of adnominals in the case of gender mismatch; context, pattern, and animacy in the case of paucal constructions. The model fitting procedure was implemented in R (R Core Development Team, 2015). The goodness-of-fit can be estimated by the concordance index which for the three models was 0.7, which is considered to be acceptable (Hosmer and Lemeshow, 2000).

For all acceptability judgment experiments we also followed the same data analysis procedure. First, the raw judgments were *z-score* transformed in order to eliminate any potential scale bias resulting from differences in how each individual interpreted the scale (Schütze and Sprouse, 2013). All the reported analyses were run on both raw and transformed data; however, there were no differences in the results. In the results reported below, we provide the transformed data. For each experiment, the results of the study were entered in a Repeated Measures ANOVA with acceptability score and {STEM TYPE, CASE} for nominalizations, {ADNOMINALS, AGREEMENT PATTERN} for gender mismatch, {CONTEXT, PATTERN, CASE} for paucal constructions as factors.

[7]As proposed by Featherston (2007), asking about the *naturalness* of the stimuli avoids reference to the informant's own production and encourages her to focus on spoken rather than written form. However, we added the opposition "good"/"bad" as a more traditional option in experimental syntax and in our opinion one which is more intuitive for naïve respondents.

[8]Training sentences included both grammatical and ungrammatical sentences, which were pre-tested on a sample of 15 respondents. In order to assess their competence at completing the task, we divided the Likert scale into two halves and checked whether respondents attributed fillers to the correct half of the scale, positive or negative. In case there were mistakes in judging the training sentences, all results from the respondent in question were excluded from the analysis.

[9]For descriptive purposes, the extraneous variables are indicated when counting the number of conditions. Different scholarly traditions treat extraneous variables differently: e.g., some may omit them when counting the final number of conditions. We suppose that including them makes the description of the stimuli more transparent. These variables were tested to ensure that they were not confounders.

[10]Given the declared number of controlled variables, the reviewer suggested that we should address the issue of statistical power. To our knowledge, the only paper that has systematically examined the statistical power of linguistic experiments is (Sprouse and Almeida, 2017). Importantly, Sprouse and Almeida (2017) estimated the sample size requirements for obtaining 80% power for a Likert scale judgment survey. This was calculated for the lowest bound for power, as only one item per experimental condition was used. The calculations were performed using resampling simulations and information about the effect size for the phenomena taken from the previous experimental study by Sprouse et al. (2013), in which judgments were tested for a random sample of 150 phenomena from Linguistic Inquiry 2001–2010. The median effect size for that sample was a *Cohen's d* of 1.61, which was taken as the "average effect size" in linguistic studies. Given this average effect size and the lowest bound for power, it was shown that ten participants provide 80% power for Likert scale experiments, i.e., in 80% of such cases acceptability rating differences can be detected at statistical significance. The experiments conducted in the present study significantly exceed the requirements stated in Sprouse and Almeida (2017).

TABLE 4 | Quantitative properties of stimuli in the experiments.

Experiment	Method	Controlled variables	Number of levels	Conditions	Sentences per condition	Target sets	Fillers
Nominalizations	Production	Type of nominalized stem	4	4	4	16	32
	Judgments	Type of nominalized stem	4	8	4	32	32
		Case of external argument	2				
Gender mismatch	Production	Adnominals in NP	8	8	2	16	32
	Judgments	Adnominals in NP	8	16	2	32	32
		Agreement pattern	2				
Paucal constructions	Production	Context	3	12	2	24	48
		Pattern	2				
		Animacy	2				
	Judgments	Context	3	24	2	48	48
		Pattern	2				
		Animacy	2				
		Case	2				

Results

In the case of production, the logistic regression models showed significant effects and interactions of the factors mentioned in 4.4 ($p < 0.001$), except animacy in the paucal construction experiment and adnominal combinations in the gender mismatch experiment. As for acceptability judgments, the ANOVA analysis revealed the following results. In the nominalization experiment there was a significant effect of STEM TYPE ($p < 0.001$) on acceptability ratings and interaction between STEM TYPE and CASE ($p < 0.001$); in the gender mismatch experiment, we found a significant effect of PATTERN ($p < 0.001$) on acceptability ratings; in the paucal construction experiment we observed significant effects of CONTEXT ($p < 0.001$), PATTERN ($p < 0.001$), and CASE ($p < 0.001$), and a significant CONTEXT-CASE interaction ($p < 0.001$). All statistical tests were run in the R environment.

As the hypothesis of our study concerns the connection between frequency of occurrence and acceptability judgments, for each phenomenon we shall review the results of both experiments jointly.

In the **nominalization** production experiment, both GEN and INSTR were available as case marking strategies for transitive stems with lexical government. With unergatives speakers only made use of GEN. As predicted by previous research, there was no variation in the control conditions: only INSTR was available for transitives, and with unaccusatives INSTR was rarely used (1% of answers). For transitive stems with lexical government GEN was more frequent, which aligns with the results from Pereltsvaig et al. (2018). The distribution of GEN and INSTR for different stems is presented in **Figure 1**.

In the nominalization judgment experiment, we observed a significant difference in acceptability rates for INSTR for different stems (**Figure 2**). Importantly, INSTR was significantly more acceptable with stems with lexical government than with unaccusative stems, baseline condition (Student's *t*-test, $p = 0.03$). The acceptability scores for unergative stems did not differ significantly from the scores for unaccusative stems. That is, both production and judgment experiments contradict the suggestion that unergatives group with transitive stems with lexical government.

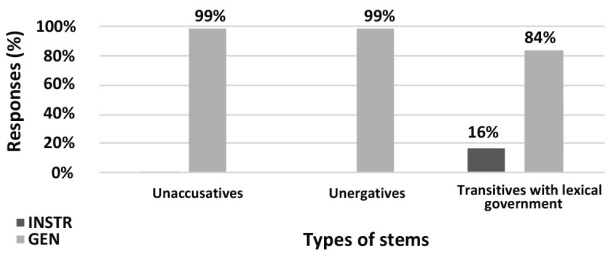


FIGURE 1 | Production experiment frequencies for nominalizations. Diagram comparing the production frequencies for INSTR (dark) and GEN (light) case marking of external argument with different nominalization stems.

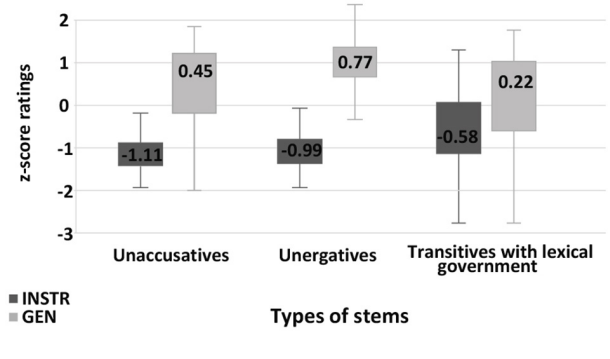


FIGURE 2 | Acceptability rating for nominalizations. Box-plot comparing the ratings for INSTR (dark) and GEN (light) case marking of external argument with different nominalization stems. Error bars represent standard deviation. Acceptability ratings are *z*-score transformed.

The **gender mismatch** production experiment showed that neither frequency nor judgment of patterns differ significantly for different combinations of adnominals. The most important result is that REFERENTIAL AGREEMENT was the most frequent pattern for all combinations of adjective modifiers, which supports the

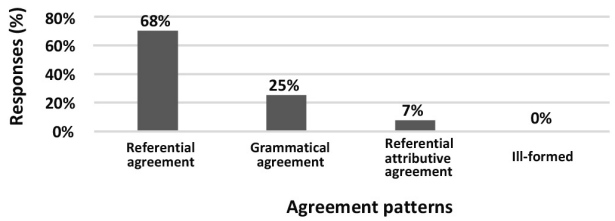


FIGURE 3 | Production experiment frequencies for gender mismatch patterns. Diagram comparing the production frequencies for different agreement patterns in gender mismatch constructions.

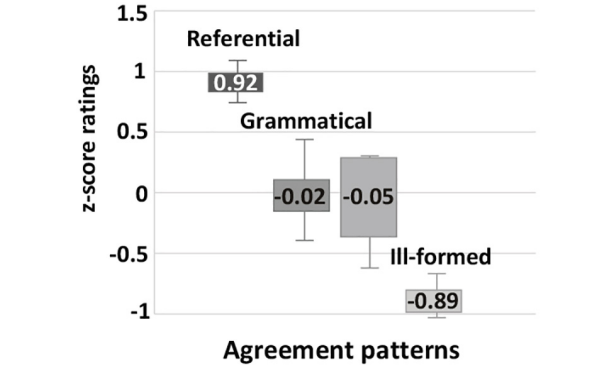


FIGURE 4 | Acceptability rating for gender mismatch patterns. Box-plot comparing the ratings for different agreement patterns in gender mismatch constructions. Error bars represent standard deviation. Acceptability ratings are *z*-score transformed.

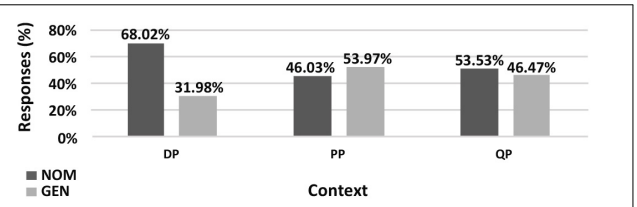


FIGURE 5 | Production experiment frequencies for paucal constructions with nominalized adjectives. Diagram comparing the production frequencies for NOM (dark) and GEN (light) case marking of nominalized adjectives in paucal constructions in different contexts.

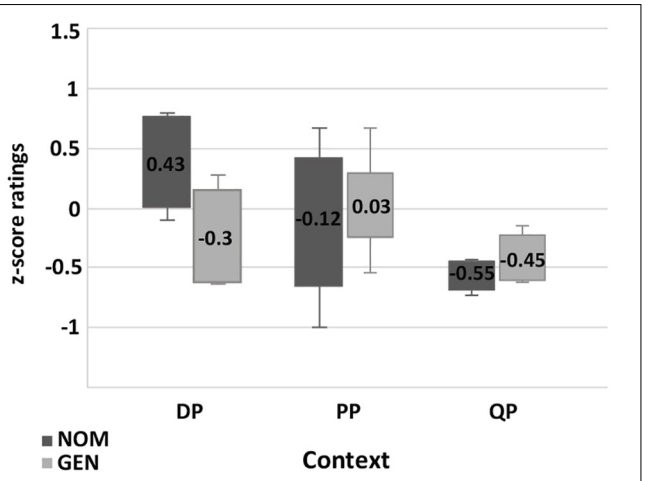


FIGURE 6 | Acceptability rating for paucal constructions with nominalized adjectives. Box-plot comparing the ratings for NOM (dark) and GEN (light) case marking of nominalized adjectives in paucal constructions in different contexts. Error bars represent standard deviation. Acceptability ratings are *z*-score transformed.

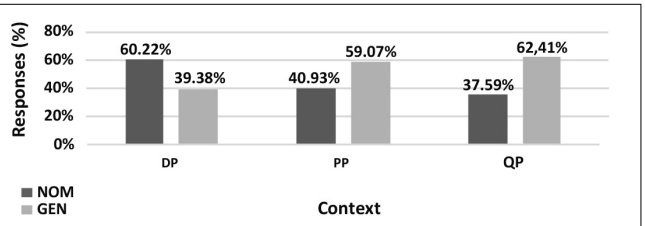


FIGURE 7 | Production experiment frequencies for paucal constructions with adjectives. Diagram comparing the production frequencies for NOM (dark) and GEN (light) case marking of adjectives in paucal constructions in different contexts.

observations of both prescriptive grammars and formal research papers (**Figure 3**). The REFERENTIAL AGREEMENT pattern was also considered the most acceptable one in the acceptability judgment experiment. It was rated significantly more acceptable than GRAMMATICAL AGREEMENT and FEMININE ATTRIBUTIVE AGREEMENT (Student's *t*-test, $p < 0.01$) (**Figure 4**).

The differences between the results of the two experiments appear when comparing the GRAMMATICAL AGREEMENT pattern and the FEMININE ATTRIBUTIVE AGREEMENT pattern. Although GRAMMATICAL AGREEMENT and FEMININE ATTRIBUTIVE AGREEMENT had significantly different frequencies in the production experiment (25% vs. 7%), they had statistically equal acceptability scores (raw means 2.92 vs. 2.75 and *z*-score means −0.02 vs. −0.05; Student's *t*-test, $p > 0.1$).

The results of the **paucal construction** experiments generally supported the hypotheses and observations reported in the previous literature. However, there are differences between the results for nominalized adjectives and those for adjectives modifying feminine nouns. In particular, for nominalized adjectives in argumental (DP) position NOM is preferred over GEN (χ^2, $p < 0.01$), while in quantificational positions (PP and QP) both NOM and GEN are available, see **Figure 5**. For attributive adjectives in argumental (DP) position NOM is preferred over GEN (χ^2, $p < 0.01$), and for attributive adjectives in quantificational positions (PP and QP) GEN is preferred over NOM (χ^2, $p < 0.01$) (**Figure 7**).

In the judgment experiment for both nominalized adjectives and adjectives that modify feminine nouns in argumental (DP) position NOM is rated as significantly more acceptable than GEN (Student's *t*-test, $p < 0.01$) (**Figures 6**, **8**). For both types of adjectives in quantificational contexts (PP and QP) NOM and GEN have almost the same acceptability ratings (Student's *t*-test, $p > 0.1$). This means that the judgment results support the

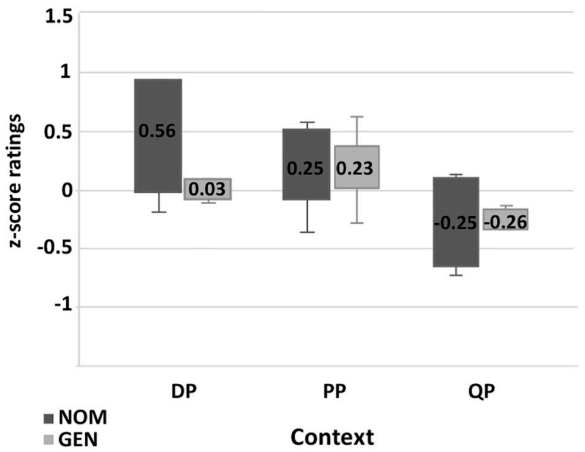


FIGURE 8 | Acceptability rating for paucal constructions with adjectives. Box-plot comparing the ratings for NOM (dark) and GEN (light) case marking of adjectives in paucal constructions in different contexts. Error bars represent standard deviation. Acceptability ratings are *z*-score transformed.

production results in all the conditions except for attributive adjectives in quantificational contexts (PP and QP). In the latter case GEN is clearly preferred in production, but NOM and GEN have almost the same acceptability ratings.

DISCUSSION OF THE EXPERIMENTAL RESULTS

As can be seen from the data analysis, the tendencies predicted theoretically are supported by the experimental data in both types of experiments. The results indicate that (i) for nominalizations derived from transitive stems with lexical government GEN is more frequent and more acceptable than INSTR, (ii) the REFERENTIAL AGREEMENT pattern is the most frequent and the most acceptable choice for gender mismatch nouns, (iii) in paucal constructions in argumental (DP) position NOM is more frequently used and is rated as more acceptable than GEN, and in paucal constructions in quantificational positions (PP and QP) NOM and GEN are both available and rated equally acceptable. However, it is worth noticing that there is no ceiling effect for any variant in the target conditions in either of the surveys.

The crucial observation is that the results of the two experiments do not necessarily coincide. In the case of gender mismatch, the two agreement patterns, GRAMMATICAL AGREEMENT and REFERENTIAL ATTRIBUTIVE AGREEMENT, are produced and rated at different levels. Similar disparities are observed in the paucal construction experiments: in quantificational context condition for adjectives there is no preference in judgments but a clear preference for GEN in usage.

The goal of our study, however, is to analyze the consistency of individual speakers over the production and perception domains. In the next section, we aim to explore whether the speakers' evaluation of the acceptability of the alternatives is consistent with the grouping based on their actual usage in production.

Analysis of the Consistency of Respondents

Adopting the view that grammar is probabilistic in nature presumes that frequencies of occurrence and acceptability scores are functions of the same grammatical constraints. The two domains are clearly non-identical, and the differences between the two modalities inevitably add noise and distortion to how the grammatical constraints are implemented. Hence, we assume that it is unreasonable to relate either the absolute or the relative size of differences in ratings to frequency differences. Instead, we suggest analyzing *relative directional differences*, viz. whether the *direction* of acceptability is predicted by production or vice versa. In case a respondent is consistent over pairs of experiments, we expect that in both production and judgments there will either be a preference toward one of the variants or both variants will be permitted and judged acceptable.

To measure the consistency of individual respondents, we checked whether each respondent who participated in both experiments rated the variant that she used in the production experiment as more acceptable than the alternative. In particular, we developed a metric which was computed as follows. For nominalizations and paucal constructions we registered (i) what the respondent produced, whether one or both alternatives, in a certain condition and (ii) which of the two alternatives the respondent rated as more acceptable in the very same condition. For the latter, we compared the mean values in raw format. Those cases where the mean values were equal were counted as if both variants were acceptable. The gender mismatch experiments were different from the two other sets in that they offered a choice of four major patterns. Nevertheless, there were very few cases where a respondent rated more than two patterns as equally acceptable, so there was no need to compute the metric for this phenomenon differently.

With the new metric we compared the same conditions across experiments that were conducted using different methodologies. As we are interested in comparing production and perception for the phenomena prone to variation, when making the comparison we took into account only those conditions that allowed for variation. Notice that the metric does not consider the same

TABLE 5 | Relative directional difference for the three experiments.

Three strategies of choice and rating	Nominalizations	Gender mismatch	Paucal constructions
1. What is produced is rated as most acceptable	55%	57%	39%
2. One alternative in one experiment, and both in the other	29%	30%	37%
2a. Both variants in production	25%	14%	23%
2b. Both variants in judgments	4%	16%	14%
3. Different alternatives in each experiment	16%	13%	24%

In each cell of the table we present the percentage of cases when the respondent demonstrated the given behavior toward an experimental condition. All conditions were taken from the production experiments.

lexical variants, as the two types of experiments contained different numbers of stimuli.

The results of the consistency analysis are presented in **Table 5**. The most striking result to emerge from the data is that, on average, respondents stick to one variant in only half of the conditions that allow for variation. For instance, for the nominalizations the metric shows that in 55% of cases the answer provided to a given condition was the same in both experiments, while in 29% of cases respondents allowed both variants in one experiment but preferred only one variant in the other, and in 16% of cases the variant produced was in fact rated as the least acceptable. The figures are even more revealing for paucal constructions. Here, the production and the choice that was rated as the most acceptable coincided in only 39% of cases, while in 37% of cases both variants were allowed in one of the experiments and only one variant in the other. In 24% of cases the variant used was rated as the least acceptable.

In the gender mismatch experiments, the preference for a single pattern was preserved in 57% of answers[11]. The gender mismatch experiments were different from the two other sets in that there was a choice to be made between four major patterns. Nevertheless, there were very few instances when a respondent rated more than two patterns as equally (highly) acceptable. In 30% the results partly coincided, with respondents showing more flexibility in one of the experiments than in the other. Finally, in 13% of answers respondents were inconsistent.

The consistency analysis also shows that in the nominalization and paucal construction experiments respondents were more likely to use both variants in production than in their acceptability judgments. For gender mismatch experiments these rates did not differ.

As there were several experimental items for one condition, we estimated whether respondents were more consistent within one condition in production or acceptability judgment experiments. The results of the computations are presented in **Table 6**. Within each experiment, we analyzed whether a respondent was consistent across different lexicalizations of a single condition.

The analysis shows that in the nominalization and paucal construction experiments there was more variability within production than in acceptability judgments. In gender mismatch experiments, there was no difference in variability. Taken together the two metrics indicate paucal constructions to be more unstable than the other two phenomena: there was much more variability in the answers given in relation to paucal constructions in both production and acceptability judgment experiments.

GENERAL DISCUSSION

The main goal of this study was to investigate how grammatical options can be distributed in the production and perception domains of a single speaker. Specifically, we hypothesized that if grammatical knowledge is indeed probabilistic, a single speaker would be consistent across the two domains of speech, providing data that follows the same grammatical constraints in both offline production and offline perception. The stated objective determined the methodology for the study: in this paper, we reported two series of experiments which involved both production and acceptability judgments. The experimental materials were based on three types of constructions in Russian which display a certain degree of variability.

Three findings from the experiments reported above can be identified as the most important. First, the experimental data in general supports the idea of alignment between acceptability ratings and frequency of occurrence. In all three pairs of experiments, the most frequent variant coincided with the one that received the highest acceptability score (GEN for transitive nominalizations with lexical government, REFERENTIAL AGREEMENT for gender mismatch nouns, NOM for paucal constructions in argumental position). Second, the results of production experiments do not always correspond to the associated acceptability ratings, even when production and ratings are provided by the same respondents. This is the case for GRAMMATICAL AGREEMENT and REFERENTIAL ATTRIBUTIVE AGREEMENT with gender mismatch nouns and for the distribution of NOM and GEN in paucal constructions in quantificational position. Third, speakers are not consistent in choosing one variant across the two types of experiment:

[11] In the gender mismatch experiment, there was more than one theoretically possible number of alternatives. Therefore, there was a methodological question concerning which situation should be recorded under the heading *usage/acceptance of both variants*: on one hand, it was important to maintain the application of a similar metric across the three sets of experiments, but on the other hand, it would be practically impossible for a respondent to use all three possible patterns within one experiment. Therefore, we recorded that *both variants* were used in a production experiment or *both variants* were rated acceptable when at least two possible variants out of three were used or rated acceptable, respectively. The properties of the gender mismatch experiment also dictate a different definition for what situations should be considered as *preference for different alternatives*. If a respondent used the REFERENTIAL AGREEMENT pattern in the production experiment but rated GRAMMATICAL AGREEMENT and REFERENTIAL ATTRIBUTIVE AGREEMENT as equally acceptable and more acceptable than REFERENTIAL AGREEMENT, this situation was counted as *preference for different alternatives* in each experiment (even though the choice was not binary).

TABLE 6 | Consistency of respondents within one experiment with respect to one condition.

	Nominalizations		Gender mismatch		Paucal constructions	
	Production	**Judgments**	**Production**	**Judgments**	**Production**	**Judgments**
The same variant within one condition (is produced/rated as most acceptable)	73%	94%	85%	82%	71%	80%
Different variants within one condition (are produced/rated as most acceptable)	27%	6%	15%	18%	29%	20%

more variation is allowed in production experiments. Moreover, variation can be characterized from the point of view of speaker consistency: different phenomena exhibit different values for consistency measures.

Inconsistency and the Diachronic Status of a Phenomenon

In this section, we would like to discuss the possible sources of inconsistency across the experiments. A plausible reason for inconsistency is the nature of the phenomena examined. As we are discussing variation, we are entering supposedly unstable language domains and examining constructions undergoing change. This change is to a great extent driven by the Economy Principle [also known as the Principle of Least Effort (Zipf, 1949)], viz., the tendency to economize on cognitive resources when conveying a message. In the context of historical linguistics, the Economy Principle is regarded as a trigger for grammatical change, since it is not economical to expend resources on several competing variants. As the existence of several options is not in accordance with expending less effort, it is expected either for the alternation to disappear (via the disappearance of one of the variants) or for the distribution of the variants to become fixed. Unless this state is achieved, we are observing different stages of language development. The periphery of variation, viz. those variants that are at the low end of the frequency spectrum, might indeed be (i) the residual effects of language evolution or, conversely, (ii) prerequisites for future changes. That is, inconsistency across the answers given by a single respondent in this case can be expected. What is remarkable is that the types of inconsistency observed differ, which means that the variation can be further characterized from this point of view.

In particular, for **nominalizations** INSTR case marking is reported as a rather new strategy (Pereltsvaig et al., 2018). This diachronic property serves as an explanation for the low frequency counts displayed by this variant. We suggest that due to its innovative nature the strategy is still rated as somewhat unacceptable even by those respondents who use it.

In cases of **gender mismatch**, REFERENTIAL AGREEMENT has been reported as the principal strategy since the 1970s (Muchnik, 1971; Crockett, 1976; Shvedova, 1980). However, while in production speakers predominantly follow a certain pattern, they also produce structurally possible alternatives to which they give equal scores: GRAMMATICAL AGREEMENT is still more frequent than FEMININE ATTRIBUTIVE AGREEMENT, but both variants have the same, rather low, level of acceptability. That is, the two alternative variants on the periphery are equalized when consciously considered. We hypothesize that these judgments reflect a gradual decrease in production frequency of the GRAMMATICAL AGREEMENT pattern in comparison to the favored REFERENTIAL AGREEMENT pattern[12].

A conceptually similar situation is found for QP contexts in **paucal constructions:** in production respondents prefer one variant, but they rate both possibilities equally when perceiving them. That is, while there is a clear leader in production, judgments reveal this only partly, via the dispersion of possible answers, which is higher for the less common variant.

We suppose that the degree of coherence of the two experiments corresponds to different stages of the evolution of the variation involved. What we observe in case of gender mismatch might be the effects of the disappearance of variation. In contrast, in the case of nominalizations we see the ongoing development of a competing variant. In the case of paucal constructions, we do not have enough diachronic data to predict the direction of change; however, Economy Principle considerations suggest that variants are becoming more fixed with respect to the structural position they take up.

Inconsistency and the Experimental Methodology

The data shows that elicited production and acceptability judgments differ with respect to how they reveal variation in language. We suggest that this inconsistency is partly dictated by the properties of the methodology used. Acceptability judgments in general show less variability. The restrictive quality of the method is revealed when analyzing whether respondents are consistent within one condition in separate experiments: within one condition the same variant is chosen more often in the judgment experiment than in the production experiment (**Table 6**). The question is what mechanisms behind the experimental methods involving production and perception determine the differences in the results.

As stated by Schütze (1996), an acceptability judgment is a reported perception of acceptability. It is not clear what the mechanisms are that help to bring about this percept: whether it is accumulated during the process of perceiving the sentence, while the respondent is comparing the actual percept with her expectations [e.g., as in the theory of forward action modeling by Garrod and Pickering (2013)], or whether the procedure is more complicated. Regardless of the specific percept model, we suppose that what is present in the case of judgment, and lacking in production, is reference to previous metalinguistic experience when deciding on an exact rating. We hypothesize that during the acceptability judgment experiment the respondent is referring to her previous experience, i.e., to the percepts of other sentences that she has perceived. Our idea is that this reference in itself produces a cognitive load that restricts the availability of the less activated elements. That is why this additional step leads to greater restrictiveness in comparison to production results.

Although the rating task makes the choice more restricted, we argue that the production method should not be generally preferred as more sensitive. Neither production nor judgment data provide direct access to the grammar: they add distortion of different kinds, as different sets of cognitive systems are involved

[12]Muchnik (1971) reported results from a questionnaire completed by 3,780 Russian native speakers, which showed that GRAMMATICAL AGREEMENT was chosen in 38.6% of cases, while in the current study GRAMMATICAL AGREEMENT was used 25.21% of the time. Although this is in line with our hypothesis, it may not be appropriate to compare the results of the earlier study with those produced by the current research due to differences in their design. Muchnik's questionnaire included only two lexical variants of the combination "high adjective + noun" and two lexical variants of the combination "noun + verb," and the questions were of the following type: *How would you say, with reference to a woman: "nice-*M *doctor" or "nice-*F *doctor"?* No context was provided for the noun phrase and the verb, and the participants' attention was drawn toward the agreement properties, which significantly lowers the ecological validity of the survey.

in the processes of production and perception. We suggest that the two experimental methods used in this paper are sensitive to different aspects of language phenomena. In particular, elicited production is better in revealing deviations from the patterns prescribed in grammars, while acceptability judgments are better at investigating to what extent a grammatical innovation has become established in the language. The combination of production and judgment data thus allows us to estimate the directionality of ongoing changes in variability and gain access to the full distribution of variants.

This observation leads to another question, namely, how the results obtained in this paper can be extrapolated to other language phenomena that do not exhibit such variability. In this study, we examined three types of construction reported in the previous literature as involving variation. However, we doubt that one can ultimately tell where the variability ends. It might be impossible to eliminate variability and ascertain whether a phenomenon is "stable" in advance of carrying out research on it. We believe that any language phenomenon should be analyzed taking into consideration both production and judgment data, as it is potentially subject to variation of the type investigated here.

Implications for Methodology

The way experimental methods are applied traditionally presupposes analyzing the sample as a whole and averaging out the individual differences. However, the properties of individual behavior toward a certain phenomenon might provide a glimpse of its current state.

Similar ideas are being developed in the field of research on bilingualism. The multidimensional concept of bilingualism cannot be treated as a categorical variable because bilingual experience shapes the way executive functioning is performed (Takahesu et al., 2018; De Bruin, 2019). Both production and comprehension processes adapt to the demands determined by the bilinguals' previous language experience: for instance, Beatty-Martínez and Dussias (2017) provide supporting evidence analyzing how individuals' production choices correlate with their code-switching strategies. The differences in how production and comprehension processes are tuned might be defined not only by language experience, but also by a set of individual-level skills such as word-decoding, working memory, and susceptibility to memory interference (Fricke et al., 2019). Rather than treating interspeaker variation as noise, an increasing number of studies propose that interspeaker variation could shed light on the linguistic architecture and how it is coordinated with other cognitive systems.

Remarkably, the results of our study suggest that linguistics could benefit from implementing an even more fine-grained approach and taking into account the behavior of each individual speaker. Differences in linguistic experience and cognitive skills supposedly should not influence the link between the production and perception domains of a single speaker. Even though respondents differ in experience with respect to certain language phenomena (e.g., poor input), we expect them to be consistent in their individual preferences across different tasks.

As the result of our study, we have devised a metric that allows us to estimate the consistency of respondents with respect to a language phenomenon in the two language domains. In particular, we have shown that inconsistency rates are far from being random, both within a single experiment and across experiments conducted with different methodologies. Importantly, the metric allows us to characterize each condition in the experiment in terms of speakers' consistency in using a certain variant. Consequently, it can also be used within a single phenomenon for a comparative analysis of conditions. We believe that the elaborated metric can be used as a formal instrument for the description of variation and will be beneficial in studying language phenomena displaying variability. Further work is needed to investigate how far speakers can be inconsistent in the production and perception of a certain phenomenon such that the phenomenon may still be considered a part of the language system. Another interesting issue is how changes in the consistency of certain individuals' behavior may influence the dynamics behind innovations in a language community. We leave these questions to future research.

CONCLUSION

The present study investigated the correspondence between offline production and offline perception in the speech of individual speakers. In our study we focused on variation and examined three types of construction that display a certain degree of variability. As can be seen from the results, using just one experimental technique would somewhat limit our understanding of the phenomena under investigation. Our data suggest that there is a correspondence between frequency of occurrence and acceptability rates. However, this correspondence is more complicated than has been stated in previous studies: different phenomena involving variation deviate from the ideal correspondence to different extents. We have shown that the combination of two sources of data provides a fuller description for cases of intralingual variation than the use of a single method. The way the data sources conform allows us to distinguish different types of variation and define unstable language domains, and, furthermore, it can serve as an additional descriptive measure.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The Commission for Ethics of Pushkin State Russian Language Institute confirmed that in accordance with the Legislation of the Russian Federation appropriate informed consent was obtained from each research participant or participant's legal guardian/next of kin, the data storage was organized in accordance with the law, and the research design was in accordance with the relevant ethical standards.

AUTHOR CONTRIBUTIONS

AG and EL have made substantial, direct and intellectual contribution to the conception and design of the work, interpretation of data, analysis, and have both approved it for publication.

FUNDING

The research underlying this article was supported by the Russian Science Foundation project #18-18-00462 "Communicative-syntactic interface: typology and grammar" at the Pushkin State Russian Language Institute.

ACKNOWLEDGMENTS

We would like to thank all the respondents who contributed their time and linguistic expertise to our experiments. We thank the editor and the reviewers for their help in improving the manuscript. All errors remain our responsibility.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00348/full#supplementary-material

REFERENCES

Adger, D., and Svenonius, P. (2011). "Features in minimalist syntax," in *The Oxford Handbook of Linguistic Minimalism*, ed. C. Boeckx (Oxford: OUP Oxford), 27–51.

Adli, A. (2011). "On the relation between acceptability and frequency," in *The Development of grammar: Language Acquisition and Diachronic Change. In Honour of JÜRGEN M. Meisel*, ed. J. M. Meisel (Amsterdam: John Benjamins Publishing), 383–404.

Asarina, A. (2009). "Gender and Adjective Agreement in Russian." in *Paper, SLS4 Annual Meeting*. Available at: http://web.mit.edu/alya/www/sls4-handout.pdf (accessed September 4, 2009).

Bailey, C.-J. (1973). *Variation and Linguistic Theory*. Arlington, VA: Center for Applied Linguistics.

Beatty-Martínez, A., and Dussias, P. (2017). Bilingual experience shapes language processing: evidence from codeswitching. *J. Mem. Lang.* 95, 173–189. doi: 10.1016/j.jml.2017.04.002

Bermel, N., Knittl, L., and Russell, J. (2018). Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells. *Corpus Linguist. Linguist. Theory* 14, 197–231. doi: 10.1515/cllt-2016-3–320032

Bickerton, D. (1975). *Dynamics of a Creole System*. Cambridge: Cambridge University Press.

Bresnan, J. (2007). "Is syntactic knowledge probabilistic? Experiments with the English dative alternation," in *Roots: Linguistics in Search of its Evidential base, 96*, eds S. Featherston, and W. Sternefeld (Berlin: Walter de Gruyter), 77–96.

Crockett, D. (1976). *Agreement in Contemporary Standard Russian*. Cambridge, MA: MIT Press.

De Bruin, A. (2019). Not all bilinguals are the same: a call for more detailed assessments and descriptions of bilingual experiences. *Behav. Sci.* 9, 33. doi: 10.3390/bs9030033

Divjak, D. (2017). The role of lexical frequency in the acceptability of syntactic variants: evidence from that clauses in Polish. *Cogn. Sci.* 4, 354–382. doi: 10.1111/cogs.12335

Featherston, S. (2007). Data in generative grammar: the stick and the carrot. *Theor. Linguis.* 33, 269–318.

Francom, J. (2009). *Experimental Syntax: Exploring The Effect of Repeated Exposure to Anomalous Syntactic Structure–Evidence From Rating and Reading Tasks*. Tucson, AZ: University of Arizona.

Fricke, M., Zirnstein, M., Navarro-Torres, C., and Kroll, J. (2019). Bilingualism reveals fundamental variation in language processing. *Biling. Lang. Cogn.* 22, 200–207. doi: 10.1017/S1366728918000482

Garrod, S., and Pickering, M. J. (2013). An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36, 329–347. doi: 10.1017/s0140525x12001495

Gerasimova, A. (2016). Parametricheskii podkhod k vnutriyazykovomu var'irovaniyu: problemy i metody [A parametric approach to intralingual variation: problems and methods]. *Rhema* 3, 63–74.

Gerasimova, A. (2019). *Variation in Agreement Features in Russian Nominals*. MA Thesis, Lomonosov Moscow State University, Moscow.

Golub, I. B. (1997). *Stilistika Russkogo Yazyka: Ucheb. Pos. [Russian stylistics: study guide]*. Moscow: Airis-press.

Graudina, L., Ickovich, V., and Katlinskaya, L. (1976). *Grammaticheskaya Pravil'nost' Russkoy Rechi (Opyt chastotno-stilisticheskogo slovarya variantov) [Grammatical correctness of Russian speech. An attempt to compiling frequency-based stylistic dictionary of variants]*. Moscow: Nauka.

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., and Snider, N. (2013). The source ambiguity problem: distinguishing the effects of grammar and processing on acceptability judgments. *Lang. Cogn. Process.* 28, 48–87. doi: 10.1080/01690965.2011.572401

Horvath, B., and Sankoff, D. (1987). Delimiting the Sydney speech community. *Lang. Soc.* 16, 179–204. doi: 10.1017/s0047404500012252

Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*. New York, NY: Wiley.

Klavan, J., and Veismann, A. (2017). Are corpus-based predictions mirrored in the preferential choices and ratings of native speakers? Predicting the alternation between the Estonian adessive case and the adposition peal 'on'. *Esuka Jeful.* 8, 59–91. doi: 10.12697/jeful.2017.8.2.03

Koptjevskaja-Tamm, M. (1993). *Nominalizations*. London: Routledge.

Kuhl, J. W. (2003). *The Idiolect, Chaos, and Language Custom Far From Equilibrium: Conversations in Morocco*. Doctoral dissertation, University of Georgia, Athens, GA.

Lau, J. H., Clark, A., and Lappin, S. (2014). "Measuring Gradience in Speakers' Grammaticality Judgements," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (Quebec City).

Lau, J. H., Clark, A., and Lappin, S. (2015). "Unsupervised prediction of acceptability judgements," in *Proceedings of the 53rd Annual Conference of the Association of Computational Linguistics*, Beijing.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cogn. Sci.* 41, 1201–1241. doi: 10.1111/cogs.12414

Levshina, N. (2015). *How to do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins Publishing Company.

Lyutikova, E. (2015). Features, agreement, and structure of the Russian noun phrase. *Russkii Yazyk v Nauchnom Osveshchenii*. 30, 44–74.

Muchnik, I. (1971). *Grammaticheskie Kategorij Glagola i Imeni v Sovremennom Russkom Literaturnom Jazyke [Verbal and Nominal Categories in Modern Standard RUSSIAN]*. Moscow: Nauka.

Pereltsvaig, A. (2006). Small Nominals. *Nat. Lang. Linguist.Theory* 24, 433–500. doi: 10.1007/s11049-005-3820-z

Pereltsvaig, A. (2015). "Nominalizations in Russian: argument structure, case, and the functional architecture of the noun phrase," in *Paper, 6th Workshop on Nominalizations*, Verona.

Pereltsvaig, A. (2017). Russian eventive nominalizations and universality of Determiner Phrase. *Rhema* 4, 108–122.

Pereltsvaig, A., Lyutikova, E., and Gerasimova, A. (2018). Case marking in Russian eventive nominalizations: inherent vs. dependent case theory. *Russian Linguist.* 37, 1–16.

Pesetsky, D. (2013). *Russian Case Morphology and the Syntactic Categories*. Cambridge, MA: MIT Press.

Phillips, C. (2009). Should we impeach armchair linguists. *Jap. Korean Linguist.* 17, 49–64.

Puškar, Z. (2017). *Hybrid Agreement: Modelling Variation, Hierarchy Effects and-Feature Mismatches*. Leipzig: Universität Leipzig.

R Core Development Team (2015). *R: A Language and Environment for Statistical Computing*. Doctoral thesis, R Foundation for Statistical Computing, Vienna.

Rothstein, R. A. (1980). "Gender and reference in Polish and Russian," in *Morphosyntax in Slavic*, eds C. V. Chvany, and R. D. Brecht (Columbus OH: Alavica), 79–97.

Schütze, C. T. (1996). *The Empirical base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Schütze, C. T., and Sprouse, J. (2013). "Judgment data," in *Research methods in linguistics*, eds R. J. Podesva, and D. Sharma (London: A&C Black), 27–50. doi: 10.1017/cbo9781139013734.004

Shkapa, M. V. (2011). Soglasovanie Opredeleniya s Sushchestvitel'nym pri Chislitel'nykh dva, tri, chetyre [Agreement Between an Adjective and a Noun Within Numerals dva, tri, chetyre]", in Problemy russkoi Stilistiki po Dannym NKRYa. Available at: https://studiorum-ruscorpora.ru/stylistics/syntax_numeral/

Shvedova, N. Y. (1980). *Russkaja Grammatika [Russian grammar]*. Moscow: AN SSSR Publication.

Sprouse, J. (2007). *A Program for Experimental syntax: Finding the Relationship Between Acceptability and Grammatical Knowledge*. College Park, MD: University of Maryland. Doctoral dissertation.

Sprouse, J. (2015). Three open questions in experimental syntax. *Linguist. Vanguard* 1, 89–100.

Sprouse, J., and Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa A J. Gen. Linguist.* 2, 14. doi: 10.5334/gjgl.236

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134, 219–248. doi: 10.1016/j.lingua.2013.07.002

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., and Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. GLOW Issue. *Linguist.Rev.* 35, 575–599. doi: 10.1515/tlr-2018-0005

Stadthagen-González, H., López, L., Parafita Couto, M. C., and Párraga, A. (2017). Using two-alternative forced choice tasks and Thurstone's law of comparative judgments for code-switching. *Linguist. Approaches Biling.* 8, 67–97. doi: 10.1075/lab.16030.sta

Steriopolo, O. (2018). Mixed gender agreement in the case of Russian hybrid nouns. *Quest. Answers Linguist.* 5, 91–106.

Svenonius, P. (2008). "the position of adjectives and other phrasal modifiers in the decomposition of DP," in *in Adjectives and Adverbs: Syntax, Semantics, and Discourse*, eds L. McNally, and C. Kennedy (Oxford: OUP), 16–42.

Takahesu, T., Emily Mech, A., and Atagi, N. (2018). Exploiting language variation to better understand the cognitive consequences of bilingualism. *Front. Psychol.* 9:1686. doi: 10.3389/fpsyg.2018.01686

Verhoeven, E., and Temme, A. (2017). "Word order acceptability and word order choice," in *Linguistic Evidence 2016 Online Proceedings*, ed. Featherston (Tübingen: Universität Tübingen.).

Wolfram, W., and Beckett, D. (2000). The role of the individual and group in earlier African American English. *Am. Speech* 75, 3–32.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Menlo Park, CA: Addison-Wesley.

Conflict of Interest: 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
published: 12 November 2020
doi: 10.3389/fpsyg.2020.513740

Modeling Human Morphological Competence

Yohei Oseki[1,2]* *and Alec Marantz*[2,3,4]

[1] Faculty of Science & Engineering, Waseda University, Tokyo, Japan, [2] Department of Linguistics, New York University, New York, NY, United States, [3] Department of Psychology, New York University, New York, NY, United States, [4] NYU Abu Dhabi Institute, New York University, Abu Dhabi, United Arab Emirates

One of the central debates in the cognitive science of language has revolved around the nature of human linguistic competence. Whether syntactic competence should be characterized by abstract hierarchical structures or reduced to surface linear strings has been actively debated, but the nature of morphological competence has been insufficiently appreciated despite the parallel question in the cognitive science literature. In this paper, in order to investigate whether morphological competence should be characterized by abstract hierarchical structures, we conducted a crowdsourced acceptability judgment experiment on morphologically complex words and evaluated five computational models of morphological competence against human acceptability judgments: Character Markov Models (Character), Syllable Markov Models (Syllable), Morpheme Markov Models (Morpheme), Hidden Markov Models (HMM), and Probabilistic Context-Free Grammars (PCFG). Our psycholinguistic experimentation and computational modeling demonstrated that "morphous" computational models with morpheme units outperformed "amorphous" computational models without morpheme units and, importantly, PCFG with hierarchical structures most accurately explained human acceptability judgments on several evaluation metrics, especially for morphologically complex words with nested morphological structures. Those results strongly suggest that human morphological competence should be characterized by abstract hierarchical structures internally generated by the grammar, not reduced to surface linear strings externally attested in large corpora.

Keywords: morphology, grammaticality, acceptability, probability, psycholinguistics, computational modeling

OPEN ACCESS

Edited by:
Viviane Marie Deprez, Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Cristiano Chesi, University Institute of Higher Studies in Pavia, Italy
Naoki Fukui, Sophia University, Japan

***Correspondence:**
Yohei Oseki
yohei.oseki@nyu.edu

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

Received: *21 November 2019*
Accepted: *22 September 2020*
Published: *12 November 2020*

Citation:
Oseki Y and Marantz A (2020) Modeling Human Morphological Competence. Front. Psychol. 11:513740. doi: 10.3389/fpsyg.2020.513740

1. INTRODUCTION

Chomsky (1957) seminally argued that the grammar categorically generates *grammatical* sentences of the language, while speakers gradiently judge *acceptable* sentences of the language, as summarized below:

> "The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not sentences of L and to study the structure of the grammatical sequences. The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones." (Chomsky, 1957, p. 13; emphasis original)

On this internalist view, syntactic competence should be characterized by abstract hierarchical structures internally generated by the grammar (Everaert et al., 2015; Ott, 2017), where

grammaticality and acceptability correspond to linguistic representation and processing, respectively, hence the familiar competence-performance distinction. The independence of the grammar from probabilities over surface linear strings was evidenced by the famous *Colorless green ideas sleep furiously* sentence, which is grammatical despite vanishingly low probabilities of linear strings (cf. Pereira, 2000; Berwick, 2018)[1].

In contrast, Lau et al. (2016) recently claimed that the grammar gradiently determines grammatical sentences of the language through probabilities of linear strings without hierarchical structures. On this externalist view, syntactic competence should be reduced to surface linear strings externally attested in large corpora, where grammaticality and acceptability are isomorphic. Specifically, computational models proposed in Natural Language Processing (NLP), such as Markov Models and Hidden Markov Models (HMMs) were trained on large corpora and evaluated against human acceptability judgments via various acceptability measures, demonstrating that probabilities of linear strings can accurately explain human acceptability judgments without hierarchical structures. In response, Sprouse et al. (2018) investigated several computational models evaluated by Lau et al. (2016) with linguistically motivated corpora and measures, and revealed that there are cost-benefit tradeoffs, where computational models accurately explained human acceptability judgments only at the expense of the categorical grammaticality distinction. That is, whether syntactic competence should be characterized by hierarchical structures or reduced to linear strings has been actively debated in the cognitive science literature.

Halle (1973) generalized the internalist view to morphology, and proposed that the grammar (i.e., word formation rules) categorically generates *grammatical* ("potential") words of the language, whereas humans gradiently judge *acceptable* ("actual") words of the language, as follows (cf. Aronoff, 1976)[2]:

> "In other words, I am proposing that the list of morphemes together with the rules of word formation define the set of *potential* words of the language. It is the filter and the information that is contained therein which turn this larger set into the smaller subset of *actual* words. This set of actually occurring words will be called the *dictionary of the language*." (Halle, 1973, p. 6; emphasis original)

Embick (2012) corroborated this internalist view of morphology, and suggested that potential words such as *confusal* have the same grammaticality status as the famous *Colorless green ideas sleep furiously* sentence, in that those words are grammatical despite never being attested in large corpora.

However, Bauer (2014) criticized the distinction between grammaticality and acceptability in morphology, and alternatively defended the externalist view of morphology with methodological emphasis on large corpora (cf. Bauer et al., 2013). Indeed, words have been traditionally treated as linear strings of morphemes without any hierarchical structures, as in finite-state models of morphology (Kaplan and Kay, 1994; Beesley and Karttunen, 2003). Moreover, there has been an implicit assumption that words are stored in the mental lexicon without any morpheme units, as in dual-route models of morphology (Pinker and Ullman, 2002) and "amorphous" models of morphology (Baayen et al., 2011).

Nevertheless, there are abundant reasons to believe that morphological competence cannot be reduced to linear strings of morphemes, with apparent differences between syntax and morphology attributed to linguistic performance (cf. Halle, 1973; Bauer, 2014): (i) recursion (e.g., *anti-missile missile*; Bar-Hillel and Shamir, 1960), (ii) center-embedding (e.g., *undeundestabilizablizeable*; Carden, 1983), (iii) long-distance dependency (e.g., *enjoyable*; Sproat, 1992), among other things. Importantly, these morphologically complex words involve nested morphological structures with both prefixes and suffixes and formally require hierarchical structures beyond linear strings (Bar-Hillel and Shamir, 1960; Langendoen, 1981; Carden, 1983). Thus, the nature of morphological competence remains to be empirically investigated.

In this paper, in order to investigate whether morphological competence should be characterized by hierarchical structures or reduced to linear strings, we conduct a crowdsourced acceptability judgment experiment on morphologically complex words and evaluate five computational models of morphological competence against human acceptability judgments. Our morphologically complex words are (i) unattested with zero surface frequencies (i.e., *potential* but not necessarily *actual* words), which increases the possibility that those words have never been encountered by participants and are thus computed from component morphemes, not retrieved from the mental lexicon (cf. Hay, 2003), and (ii) trimorphemic with linear (e.g., *digit-al-ly*) and nested (e.g., *un-predict-able*) morphological structures, the latter of which can only be modeled with hierarchical structures (cf. Libben, 2003, 2006). The computational models investigated in this paper are 1. Character Markov Models (Character) with character linear strings, 2. Syllable Markov Models (Syllable) with syllable linear strings, 3. Morpheme Markov Models (Morpheme) with morpheme linear strings, 4. Hidden Markov Models (HMM) with part-of-speech (POS) linear strings, and 5. Probabilistic Context-Free Grammars (PCFG) with hierarchical structures[3]. Moreover, those computational models are evaluated against human acceptability judgments through the acceptability measure called *syntactic log-odds ratio* (SLOR; Pauls and Klein, 2012) and the evaluation metrics including effect and deviance

[1]Due to the ill-posed relationship between grammaticality and acceptability, grammatical sentences may become unacceptable (e.g., garden-path sentences), while ungrammatical sentences can become acceptable (e.g., grammatical illusions).

[2]Halle (1973) proposed that potential words such as *confusal* are assigned the feature [– Lexical Insertion], so that those words can be generated by the grammar, but never inserted into any actual sentences of the language.

[3]Recurrent neural networks (RNNs) were also investigated in the previous literature (Lau et al., 2016; Sprouse et al., 2018), but whether RNNs can implicitly represent hierarchical structures has been intensively debated with mixed results (cf. Linzen et al., 2016; Sennhauser and Berwick, 2018). Thus, as a first approximation, we start with classic but interpretable computational models and leave state-of-the-art but uninterpretable models like RNNs for future research.

accuracies, as well as an evaluation metric called *residual accuracy* proposed here to quantify the division of labor among computational models.

This paper is organized as follows. Section 2 describes the crowdsourced acceptability judgment experiment, computational models of morphological competence, and evaluation metrics to statistically compare acceptability judgments and computational models. Section 3 presents descriptive statistics of the acceptability judgment experiment and accuracies of the computational models on several evaluation metrics. Section 4 summarizes and interprets the results in the broader theoretical context. Section 5 concludes this paper.

2. METHODS

2.1. Participants

The participants were 180 native English speakers crowdsourced on Amazon Mechanical Turk (AMT). They provided electronic informed consent and were paid $2/h for their participation. We excluded 14 participants whose native language was not reported to be English ($n = 5$) or whose birthplace was not reported to be the USA ($n = 9$), resulting in 166 participants included in the statistical analyses.

2.2. Stimuli

The stimuli were created based on the CELEX lexical database (Baayen et al., 1995). The specific stimuli creation procedure consisted of several steps. First, every word was extracted from the English morphology lemma corpus (`eml.cd`) available from the CELEX, hence 52,447 words. Second, the words with stem allomorphy ("StrucAllo"), orthographic substitution ("StrucSubst"), or semantic opacity ("StrucOpac") were excluded, hence 36,800 words. Third, morphological structures ("StrucLab") were transformed from the CELEX format (e.g., ((teach)[V], (er)[N|V.])[N]) to the Penn Treebank format (e.g., (N (V teach) er)). Fourth, the remaining words were categorized into three types ("MorphStatus"): monomorphemic words (M; $n = 7{,}401$), zero conversion words (Z; $n = 7{,}375$), and morphologically complex words (C; $n = 9{,}342$), which were further subcategorized into bimorphemic words ($n = 7{,}383$), trimorphemic linear words ($n = 1{,}668$), and trimorphemic nested words ($n = 291$). The three subcategories of morphologically complex words were defined as $[_X\ [_Y\ \sqrt{\text{Root}}]\ \text{Suffix}]$ or $[_X\ \text{Prefix}\ [_Y\ \sqrt{\text{Root}}]]$ (bimorphemic), $[_X\ [_Y\ [_Z\ \sqrt{\text{Root}}]\ \text{Suffix}]\ \text{Suffix}]$ (trimorphemic linear), and $[_X\ \text{Prefix}\ [_Y\ [_Z\ \sqrt{\text{Root}}]\ \text{Suffix}]]$ (trimorphemic nested), where prefixes are attached higher than suffixes. Fifth, trimorphemic linear and nested morphological structures were extracted from trimorphemic linear and nested words, respectively. Specifically, for each outer suffix in trimorphemic linear words ($n = 48$), the possible local combinations with inner suffixes were computed, among which the suffix-suffix combination with the highest type frequency was accepted as trimorphemic linear morphological structure if (i) type frequency ≥ 5 and (ii) the outer suffix is productive (Plag and Baayen, 2009). In the same vein, for each outer prefix in trimorphemic nested words ($n = 58$), the possible non-local combinations with inner suffixes were computed, among which the prefix-suffix combination with the highest type frequency was accepted as trimorphemic nested morphological structure if (i) type frequency ≥ 2 and (ii) the outer prefix is productive (Zirkel, 2010)[4]. This procedure resulted in 10 linear morphological structures and eight nested morphological structures, as summarized below (N = noun, V = verb, A = adjective, B = adverb):

- Linear morphological structures
  1. $[_A\ [_N\ [_V\ \sqrt{\text{ROOT}}]\ \text{ion}]\ \text{al}]$
  2. $[_N\ [_A\ [_V\ \sqrt{\text{ROOT}}]\ \text{able}]\ \text{ity}]$
  3. $[_N\ [_N\ [_V\ \sqrt{\text{ROOT}}]\ \text{or}]\ \text{ship}]$
  4. $[_N\ [_V\ [_A\ \sqrt{\text{ROOT}}]\ \text{ize}]\ \text{er}]$
  5. $[_V\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{al}]\ \text{ize}]$
  6. $[_B\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{ic}]\ \text{ally}]$
  7. $[_B\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{al}]\ \text{ly}]$
  8. $[_N\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{y}]\ \text{ness}]$
  9. $[_N\ [_N\ [_V\ \sqrt{\text{ROOT}}]\ \text{ion}]\ \text{ist}]$
  10. $[_N\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{al}]\ \text{ism}]$
- Nested morphological structures
  1. $[_N\ \text{pre}\ [_N\ [_V\ \sqrt{\text{ROOT}}]\ \text{ion}]]$
  2. $[_A\ \text{sub}\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{al}]]$
  3. $[_A\ \text{super}\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{al}]]$
  4. $[_A\ \text{inter}\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{al}]]$
  5. $[_A\ \text{over}\ [_A\ [_N\ \sqrt{\text{ROOT}}]\ \text{ous}]]$
  6. $[_N\ \text{non}\ [_N\ [_V\ \sqrt{\text{ROOT}}]\ \text{ion}]]$
  7. $[_V\ \text{de}\ [_V\ [_A\ \sqrt{\text{ROOT}}]\ \text{ize}]]$
  8. $[_A\ \text{un}\ [_A\ [_V\ \sqrt{\text{ROOT}}]\ \text{able}]]$

Finally, novel morphologically complex words were created based on the linear and nested morphological structures generated above. Specifically, for each linear morphological structure, the possible stems were extracted from the subcategory of bimorphemic words whose token frequency is ≥ 20 and whose inner suffix and syntactic category match with the linear morphological structure. For example, for the linear morphological structure $[_A\ [_N\ [_V\ \sqrt{\text{Root}}]\ \text{ion}]\ \text{al}]$, the bimorphemic word *computation* with the structure $[_N\ [_V\ \sqrt{\text{Compute}}]\ \text{ion}]$ is the possible stem. Then, one stem was randomly selected from the possible stems and inserted into the linear morphological structure with orthographic adjustments performed (if necessary), and the resultant word was accepted as a novel morphologically complex linear word if unattested in (i) the CELEX lexical database and (ii) the list of socially inappropriate words. Similarly, for each nested morphological structure, the possible stems were extracted from the subcategory of bimorphemic words whose token frequency is ≥ 20 and whose inner suffix and syntactic category match with the nested morphological structure. Then, one stem was randomly selected from the possible stems and inserted into the nested morphological structure with orthographic adjustments performed (if necessary), and the

[4]The type frequency threshold for nested morphological structures was lower than for linear morphological structures, because the trimorphemic nested words ($n = 291$) were inherently sparse relative to the trimorphemic linear words ($n = 1{,}668$).

resultant word was accepted as a novel morphologically complex nested word if unattested in (i) the CELEX lexical database and (ii) the list of socially inappropriate words. Importantly, syntactic (i.e., syntactic categories), morphological (i.e., affix combinations), and phonological (i.e., orthographic adjustments) selectional restrictions were explicitly considered, while semantic selectional restrictions were not controlled because those novel morphologically complex words are intended as potential but not actual words, such as *confusal* (Halle, 1973; Embick, 2012)[5]. This final step was repeated until 300 linear and 300 nested trimorphemic words were created, while alternating between linear and nested morphological structures, hence 600 words in total. No roots were repeated in order to avoid potential priming effects across two morphological structures, and those algorithmically generated words were also double-checked by three native English speakers[6]. The stimuli are summarized in **Table 1**.

2.3. Procedure

The 600 novel morphologically complex words were distributed into six different lists of 100 unique words (50 linear and 50 nested words). Each list was randomized and the corresponding reversed list was created, resulting in 12 different lists. Each participant ($n = 180$) was randomly assigned to one of the 12 lists, so that each list was completed by 15 different participants with fixed order. Consequently, there are 30 trials for each word (15 trials from the originally randomized list and 15 trials from the reversed list), hence 18,000 trials (600 words * 30 trials) in total. We excluded 14 participants ($n = 14 * 100 = 1{,}400$) and incomplete trials ($n = 61$), resulting in 16,539 trials included in the statistical analyses.

The experiment was an acceptability judgment paradigm administered on Amazon Mechanical Turk (AMT) and implemented in HTML, where the participants judged each novel morphologically complex word on the Likert scale from 1 ("very bad") to 7 ("very good"). In order to ensure that the same participants do not complete the same experiment more than once, the experiment was assigned a unique color code and the AMT workers were asked not to complete the experiments with the same color code more than once per day, given that the entire experiment will be completed within 1 day. Before the experiment, demographic information was collected including gender, age, native language, and birthplace. The instructions are shown below:

> "In this experiment, you will read English words, and determine whether you think they are *possible* English words. We are not concerned with whether these words are *actual* English words already listed in a dictionary. Instead, we are interested in whether these words could be used by a native speaker of English. You will rate the word on a scale from 1 (very bad) to 7 (very good). Here are two examples: one that is very bad and one that is very good."

Importantly, since several pilot experiments suggested that the participants tend to judge novel morphologically complex words based on whether they have ever seen those words before, we explicitly emphasized the contrast between *possible* and *actual* words (Halle, 1973), which encouraged the participants to process the words even if they have never encountered those words before. Then, "very good" (i.e., *teacher*) and "very bad" (i.e., *readize*) bimorphemic examples were presented to familiarize the participants with the Likert scale. Finally, after the additional instruction "There are 100 words for you to rate. You must rate all of them in order to be paid for the experiment," the experiment started where 100 words were presented with their own Likert scales on the same HTML page. The experiment was piloted with *turktools* (Erlewine and Kotek, 2016) in Python and double-checked by three native English speakers. The experiment lasted for about 10 min[7].

2.4. Computational Models

Five computational models were implemented with Natural Language Tool Kit package (Bird et al., 2009) in Python: Character Markov Model with character linear strings, Syllable Markov Model with syllable linear strings, Morpheme Markov Model with morpheme linear strings, Hidden Markov Model (HMM) with part-of-speech (POS) linear strings, and Probabilistic Context-Free Grammar (PCFG) with hierarchical structures. Those models were trained on the entire CELEX lexical database ($n = 52{,}477$) via Maximum Likelihood Estimation with token weighting and Lidstone smoothing at $\alpha = 0.1$, and evaluated against human acceptability judgments of novel morphologically complex words ($n = 600$). The architectures of Markov Model, HMM, and PCFG are summarized below.

2.4.1. Markov Model

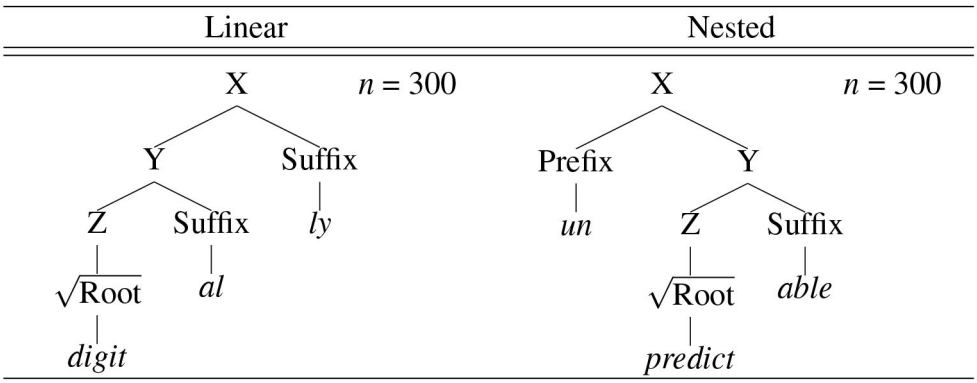
Markov Models (also called n-gram models) are defined by n-order Markov processes that compute transition probabilities of linguistic units (e.g., characters, syllables, morphemes) at position i given i–n context (e.g., $P(x_i|x_{i-n}, x_{i-1})$). Since the length of morphologically complex words is inherently limited relative to syntactically complex sentences, Markov Models were defined with $n = 1$ (i.e., bigram models), which compute transition probabilities of linguistic units at position i given the immediately

[5]Embick (2012) suggested that those potential words become acceptable if they carve out new "semantic space," which can be computationally modeled via Functional Representations of Affixes in Compositional Semantic Space (FRACCSS; Marelli and Baroni, 2015), the distributional semantic model which computes meanings of novel morphologically complex words from their component morphemes.

[6]As an anonymous reviewer suggested, the same roots in both morphological structures would help cancel out differences in specific semantic selectional restrictions between roots and inner suffixes across nested and linear morphological structures (e.g., *knowable* vs. **seeable*, as in *unknowable* vs. **seeability*). However, we prioritized not repeating roots within the experiment against controlling this semantic factor across two morphological structures.

[7]While this extended acceptability judgment paradigm might cause the participants to perform meta-linguistic (as opposed to spontaneous) judgments, we decided to adopt this design choice at the expense of spontaneous performance. In addition, the possibility that the same words were re-judged by the same participants multiple times can be safely excluded based on (i) average time per assignment (i.e., 10 min 23 s) and (ii) the incentive of AMT workers (i.e., complete as many assignments as possible).

TABLE 1 | Novel morphologically complex words unattested with zero surface frequencies and trimorphemic with linear and nested morphological structures: 300 linear words (with two inner and outer suffixes) and 300 nested words (with inner suffixes and outer prefixes), hence 600 words in total.



preceding unit (e.g., $P(x_i|x_{i-1})$)[8]. For training, Markov Models were unsupervisedly trained on character strings (Character Markov Model), syllable strings (Syllable Markov Model), and morpheme strings (Morpheme Markov Model), respectively, where character and morpheme strings were available from the CELEX lexical database, while syllable strings were generated with the `syllabify` module implemented in Python by Kyle Gorman through ARPABET transcriptions assigned by LOGIOS Lexicon Tool in the Carnegie Mellon University Pronouncing Dictionary. For testing, those trained Markov Models then computed probabilities of morphologically complex words as products of their component transition probabilities. Markov Models are sequential models, which should accurately predict local dependencies of linear words (e.g., *digitally*), but not non-local dependencies of nested words (e.g., *unpredictable*) because component local dependencies (e.g., **unpredict*) are unattested in the training data.

2.4.2. Hidden Markov Model (HMM)

HMMs generalize Markov Models with *n*-order Markov processes defined over "hidden" linear strings. HMMs compute transition probabilities of part-of-speech (POS) tags at position *i* given *i*–*n* context (e.g., $P(t_i|t_{i-n}, t_{i-1})$), and emission probabilities of morphemes at position *i* given POS tags at the same position *i* (e.g., $P(m_i|t_i)$). Like Markov Models, HMMs were also defined with $n = 1$, which compute transition probabilities of POS tags at position *i* given the immediately preceding POS tag (e.g., $P(t_i|t_{i-1})$). For training, HMMs were supervisedly trained on tagged morpheme strings generated from morphological structures available from the CELEX lexical database (e.g., [(*accident*, N), (*al*, A), (*ly*, B)]). For testing, those trained HMMs then computed probabilities of morphologically complex words as products of component transition and emission probabilities via the forward algorithm which computes the sum of path probabilities of structurally ambiguous words (Rabinar, 1989)[9]. HMMs are also sequential models, which should accurately predict local dependencies of linear words (e.g., N-A-B for *digitally*), but only approximate non-local dependencies of nested words (e.g., *unpredictable*) if component local dependencies (e.g., A-V for **unpredict*) are attested in the training data.

2.4.3. Probabilistic Context-Free Grammar (PCFG)

PCFGs generalize Context-Free Grammars (CFGs) with probability distributions defined over hierarchical structures. PCFGs compute non-terminal probabilities of right-hand sides given left-hand sides of non-terminal production rules (e.g., $P(rhs|lhs)$), and terminal probabilities of right-hand side terminals given left-hand side non-terminals of terminal production rules (e.g., $P(m_i|t_i)$), equivalent to HMM emission probabilities. Non-terminal production rules are head-lexicalized, which model syntactic selectional restrictions of derivational affixes (e.g., N → A *ness*). For training, PCFGs were supervisedly trained on morphological structures available from the CELEX lexical database (e.g., $[_B\ [_A\ [_N$ *accident*] *al*] *ly*]). For testing, those trained PCFGs then computed probabilities of morphologically complex words as products of component non-terminal and terminal probabilities via the Earley parser which computes the sum of tree probabilities of structurally ambiguous words (Earley, 1970; Stolcke, 1995)[10]. PCFGs are hierarchical models, which should accurately predict not only local dependencies of linear words (e.g., [[*digit-al*]-*ly*]), but also non-local dependencies of nested words (e.g., [*un*-[*predict-able*]]).

2.5. Statistical Analyses

Mixed-effects regression models were implemented with the `lme4` package (Bates et al., 2015) in R. The baseline regression model was first fitted with individual acceptability judgments as the dependent variable (where the acceptability judgments

[8]First-order Markov Models append one word initial symbol <w> as the necessary context to estimate transition probabilities of first morphemes.

[9]We also tested the Viterbi algorithm which computes the max of multiple paths of structurally ambiguous words, but since most probability mass was allocated to the best path, there were no substantial differences between forward and Viterbi algorithms.

[10]In the same vein, we also tested the Viterbi parser which computes the max of multiple trees of structurally ambiguous words, but since most probability mass was allocated to the best tree, there were no substantial differences between Earley and Viterbi parsers.

were z-score transformed to eliminate scale biases; Sprouse et al., 2018) and by-subject, by-word, and by-order random intercepts as random effects. Control variables, such as word length and morpheme frequency will be explained by the acceptability measure, thus not included in the baseline regression model. Then, for each computational model, the target regression model was fitted, where the acceptability measure was included as the fixed effect and random effects were held constant. Mixed-effects regression models were fitted via Maximum Likelihood Estimation with `nlminb` optimizer in `optimx` package and the maximum number of iterations R permits. Given that the baseline and target regression models are minimally different in the acceptability measure, computational models can be evaluated with nested model comparisons via log-likelihood ratio tests based on the χ^2-distribution with $df = 1$, where df is the difference in the number of parameters between two nested models.

2.6. Evaluation Metrics

2.6.1. Syntactic Log-Odds Ratio (SLOR)

The acceptability measure called *syntactic log-odds ratio* (SLOR; Pauls and Klein, 2012) is the linking hypothesis to bridge between probability estimates computed by models and acceptability judgments produced by humans (Lau et al., 2016; Sprouse et al., 2018). SLOR is defined as Equation (1):

$$SLOR = \frac{\log p_w(\zeta) - \log p_m(\zeta)}{|\zeta|} \quad (1)$$

where ζ is the morphologically complex word, $|\zeta|$ is the word length, $p_w(\zeta)$ is the word probability computed by models, and $p_m(\zeta)$ is the morpheme probability defined as $p_m(\zeta)=\prod_{m\in\zeta} p(m)$. SLOR was employed in this paper, rather than the mere correlation metric between probability and acceptability, in order to (i) control confounding factors, such as word length (i.e., $|\zeta|$) and morpheme frequency [i.e., $p_m(\zeta)$] and focus exclusively on morphological structures, and (ii) keep the evaluation procedure maximally comparable to the previous literature (Lau et al., 2016; Sprouse et al., 2018).

2.6.2. Effect Accuracy

Three evaluation metrics can be derived from SLOR based on effect sizes, deviance statistics, and residual errors. The first evaluation metric called *effect accuracy* is defined as Equation (2):

$$EA(model) = |d_{human} - d_{model}| = |\Delta d| \quad (2)$$

where d_{human} and d_{model} are Cohen's d estimated from human acceptability judgments and model SLOR scores, respectively, where Cohen's d is defined as $d = \frac{\mu_1 - \mu_2}{s}$. That is, the effect accuracy measures the absolute difference in effect sizes between human acceptability judgments and model SLOR scores, so that the lower the effect accuracy is, the more accurate the computational model is (i.e., the computational model with the effect size more comparable to the humans' is more accurate).

2.6.3. Deviance Accuracy

The second evaluation metric called *deviance accuracy* is defined as Equation (3):

$$DA(model) = D_{base} - D_{model} = \Delta D \quad (3)$$

where D_{base} and D_{model} are deviance statistics extracted from baseline and target regression models with and without model SLOR scores, respectively, where deviance statistics intuitively quantify the global error between human acceptability judgments and model SLOR scores for each computational model. That is, the deviance accuracy measures the decrease in deviance statistic from baseline to target models, so that the higher the deviance accuracy is, the more accurate the computational model is (i.e., the computational model with lower deviance statistic is more accurate).

2.6.4. Residual Accuracy

The third new evaluation metric called *residual accuracy* is proposed here as Equation (4):

$$RA(model) = \sum_{i=1}^{n} |\epsilon_{base}(w_i)| - |\epsilon_{model}(w_i)| = \sum_{i=1}^{n} \Delta|\epsilon(w_i)| \quad (4)$$

where ϵ_{base} and ϵ_{model} are residual errors extracted from baseline and target regression models with and without model SLOR scores, respectively, where residual errors intuitively quantify the local error between human acceptability judgments and model SLOR scores for each morphologically complex word. That is, the residual accuracy can measure the division of labor among computational models with respect to linear and nested morphological structures, so that the higher the residual accuracy is, the more accurate the computational model is (i.e., the computational model with lower residual error is more accurate).

3. RESULTS

3.1. Descriptive Statistics

Descriptive statistics of the acceptability judgment experiment are summarized in **Figure 1**, where the x-axis represents individual acceptability judgments z-score transformed for each participant, and the y-axis shows probability densities. Descriptive statistics are separated into linear and nested structures.

Importantly, descriptive statistics confirm that the participants were not biased toward only the upper range of the Likert scale, despite the fact that only morphologically complex words (i.e., grammatical words) were tested in this experiment without any morphologically complex nonwords (i.e., ungrammatical words). In addition, the distributions of two morphological structures seem to be bimodal as if both grammatical and ungrammatical words are included in the experiment (cf. Sprouse et al., 2018), suggesting that successful computational models should be balanced and fitted equally well to two morphological structures.

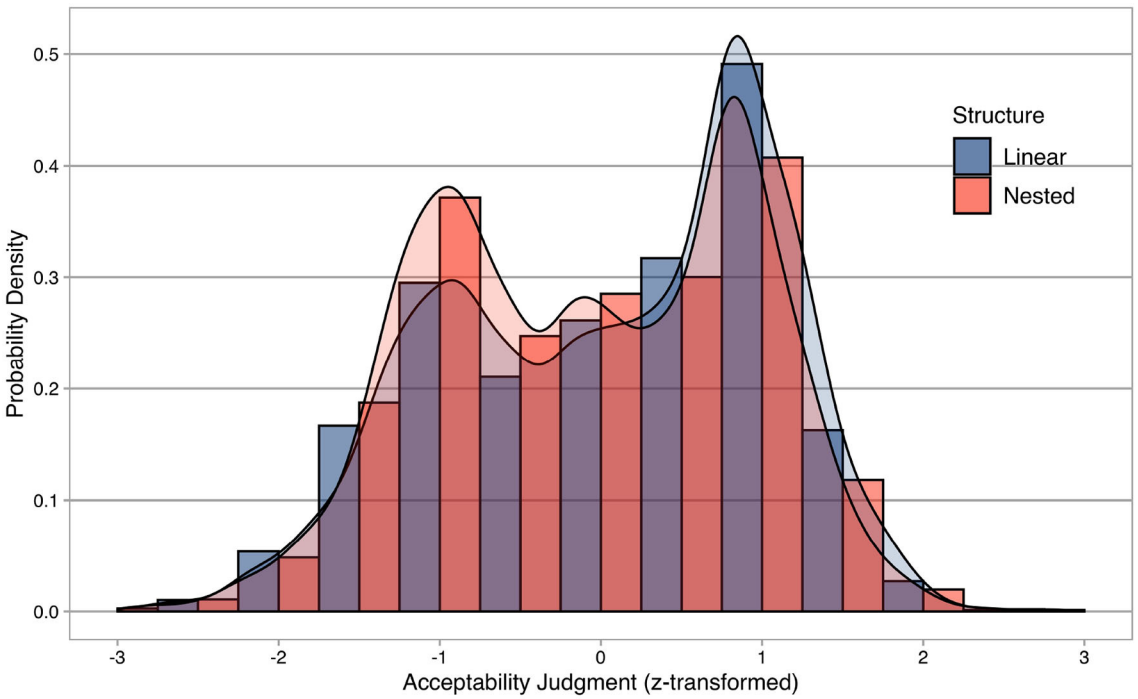


FIGURE 1 | Descriptive statistics of the acceptability judgment experiment. The *x*-axis represents individual acceptability judgments z-score transformed for each participant, while the *y*-axis shows probability densities. Descriptive statistics are separated into linear (blue) and nested (red) structures.

TABLE 2 | Effect accuracies of computational models.

Model	Linear	Nested	*t*	*p*	*d*	Δ*d*
Human	4.67	4.39	3.39	<0.001***	0.28	—
Character	−6.17	−6.31	0.63	ns	0.05	0.23
Syllable	−1.96	−2.22	0.98	ns	0.08	0.20
Morpheme	2.15	1.47	9.08	<0.001***	0.74	0.46
HMM	−0.85	−1.47	11.51	<0.001***	0.94	0.66
PCFG	1.35	1.18	2.68	<0.01**	0.22	**0.06**

*Mean acceptability judgments of linear and nested morphological structures, t-values, p-values, Cohen's d, and effect accuracies (i.e., absolute differences in Cohen's d from human acceptability judgments) are presented for each computational model; **p < 0.05, **p < 0.01, ***p < 0.001; Bold value represents best performance.*

3.2. Effect Accuracy

Effect accuracies of computational models are summarized in **Table 2**, where mean acceptability judgments of linear and nested morphological structures, *t*-values, *p*-values, Cohen's *d*, and effect accuracies (i.e., absolute differences in Cohen's *d* from human acceptability judgments) are presented for each computational model.

Independent two-sample *t*-tests indicated that the mean acceptability judgments were significantly different between linear and nested morphological structures for Human ($t = 3.39$, $p < 0.001^{***}$, $d = 0.28$), Morpheme ($t = 9.08$, $p < 0.001^{***}$, $d = 0.74$), HMM ($t = 11.51$, $p < 0.001^{***}$, $d = 0.94$), and PCFG ($t = 2.68$, $p < 0.01^{**}$, $d = 0.22$), where linear morphological structures were judged as more acceptable than nested morphological structures. Among those computational models, PCFG was most accurate with the minimal absolute difference in Cohen's *d* from human acceptability judgments ($\Delta d = 0.06$), while Morpheme and HMM were less accurate with the overestimated absolute differences in Cohen's *d* from human acceptability judgments ($\Delta d = 0.46$, $\Delta d = 0.66$), respectively.

3.3. Deviance Accuracy

Deviance accuracies of computational models are summarized in **Figure 2**, where the *x*-axis represents computational models, and the *y*-axis shows deviance accuracies (i.e., decreases in deviance statistics from the baseline model). The horizontal dashed line is $\chi^2 = 3.84$, the critical χ^2-statistic at $p = 0.05$ with $df = 1$.

Nested model comparisons revealed that the deviance statistics were significantly different between the baseline model and the target models for Morpheme ($\chi^2 = 4.55$, $p < 0.05^{*}$), HMM ($\chi^2 = 6.3$, $p < 0.05^{*}$), and PCFG ($\chi^2 = 18.04$, $p < 0.001^{***}$). Among those computational models, PCFG was most accurate with the maximal decrease in deviance statistics from the baseline model, while Morpheme and HMM were less accurate with smaller decreases in deviance statistics from the baseline model. In addition, nested model comparisons among computational models confirmed that PCFG significantly outperformed Morpheme ($\chi^2 = 13.82$, $p < 0.001^{***}$) and HMM ($\chi^2 = 11.75$, $p < 0.001^{***}$), respectively.

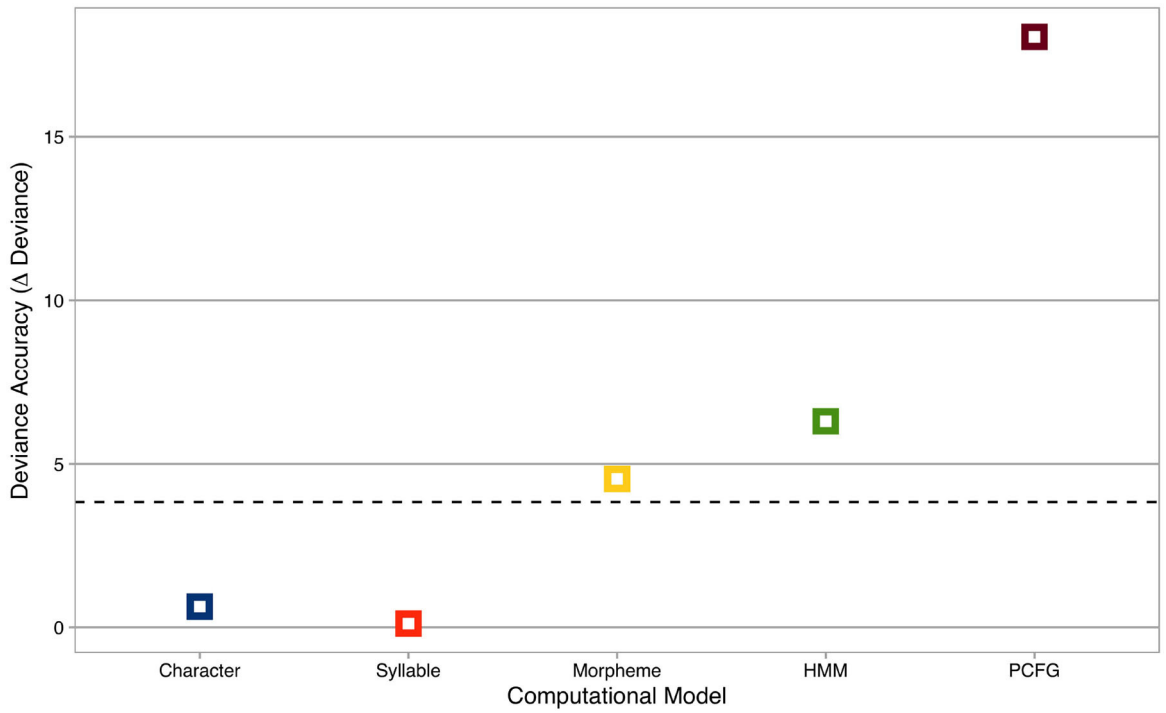


FIGURE 2 | Deviance accuracies of computational models. The *x*-axis represents computational models, while the *y*-axis shows deviance accuracies (i.e., decreases in deviance statistics from the baseline model). Colors indicate computational models: blue = Character Markov Model, orange = Syllable Markov Model, yellow = Morpheme Markov Model, green = Hidden Markov Model, brown = Probabilistic Context-Free Grammar. The horizontal dashed line is $\chi^2 = 3.84$, the critical χ^2-statistic at $p = 0.05$ with $df = 1$.

3.4. Residual Accuracy

In order to analyze and interpret the three "morphous" computational models statistically significant on deviance accuracy (i.e., Morpheme Markov Model, HMM, and PCFG), residual accuracies of computational models are summarized in **Figure 3**, where the *x*-axis represents computational models (without Character and Syllable Markov Models, which were not statistically significant on deviance accuracy), and the *y*-axis shows residual accuracies (i.e., decreases in absolute residual errors from the baseline model). Residual accuracies are categorized into linear and nested morphological structures and averaged across individual derivational affixes. The horizontal dashed line is a "tie" borderline where computational models make the same predictions as the baseline model. Positive and negative residual accuracies mean better and worse predictions relative to the baseline model, respectively.

An interesting mirror image emerged between linear and nested morphological structures. For linear morphological structures, sequential models, such as Morpheme Markov Model and HMM showed higher residual accuracies than the hierarchical model. In contrast, for nested morphological structures, the hierarchical model, namely PCFG, was relatively better than sequential models, although residual accuracies were absolutely negative for all three computational models, potentially suggesting that those computational models were overfitted to linear morphological structures and thus worsened the baseline model.

4. DISCUSSION

In summary, we have conducted a crowdsourced acceptability judgment experiment on novel morphologically complex words and then evaluated five computational models of morphological competence against human acceptability judgments via three evaluation metrics. Consequently, both effect and deviance accuracies consistently demonstrated that "morphous" computational models with morpheme units (Morpheme Markov Models, HMM, and PCFG) were more accurate than "amorphous" computational models without morpheme units (Character and Syllable Markov Models). For effect accuracies, "morphous" models correctly predicted the significant differences in effect sizes between linear and nested morphological structures like humans, while "amorphous" models underestimated the differences between those two morphological structures. In the same vein, for deviance accuracies, "morphous" models outperformed "amorphous" models which failed to even reach statistical significance relative to the baseline model. Taken together, these results strongly suggest that morphemes are psychologically real (Marantz, 2013), contrary to "amorphous" models of morphology (Baayen et al., 2011; Ackerman and Malouf, 2013).

More importantly, among those "morphous" models, the hierarchical model, namely PCFG with abstract hierarchical structures, was most accurate on both effect and deviance evaluation metrics as compared to sequential models (Morpheme

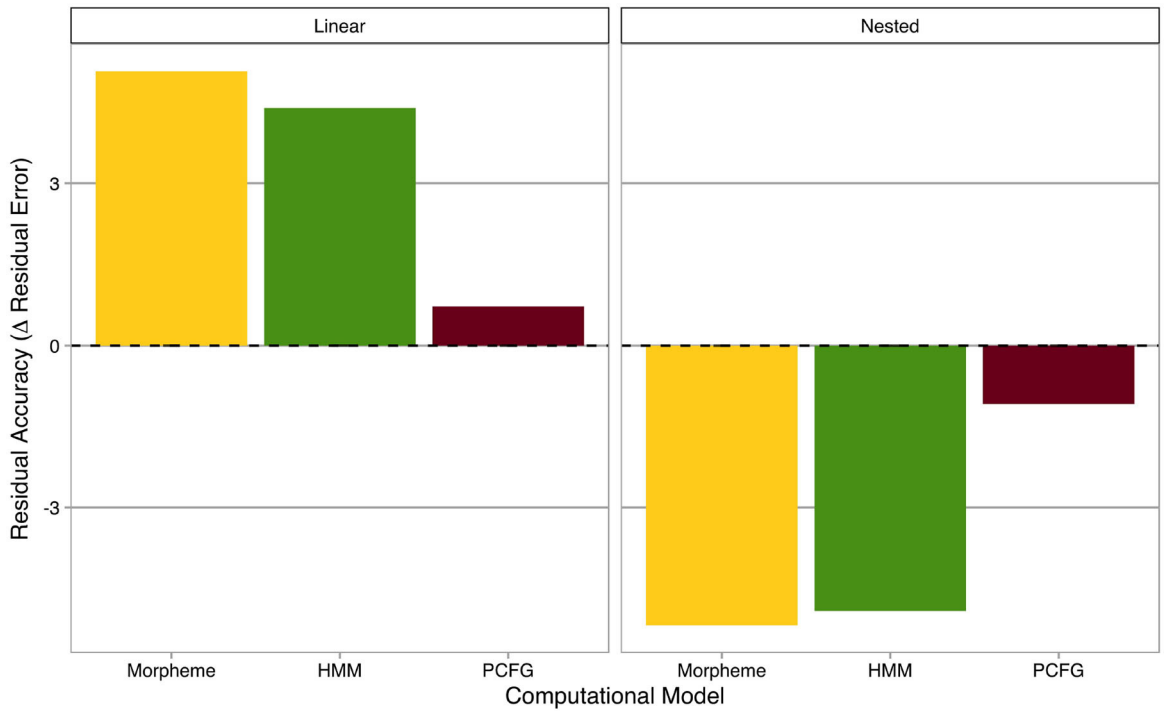


FIGURE 3 | Residual accuracies of computational models. The *x*-axis represents computational models, while the *y*-axis shows residual accuracies (i.e., decreases in absolute residual errors from the baseline model). Residual accuracies are categorized into linear (left) and nested (right) morphological structures. The horizontal dashed line is a "tie" borderline where computational models make the same predictions as the baseline model. Positive and negative residual accuracies mean better and worse predictions relative to the baseline model, respectively.

Markov Model and HMM). For effect accuracies, PCFG most accurately approximated the human effect size between linear and nested morphological structures, whereas sequential models overestimated the effect sizes between those two morphological structures. Similarly, for deviance accuracies, PCFG outperformed sequential models by a large margin. Overall, these results indicate that PCFG is the most "human-like" computational model of morphological competence, contrary to finite-state models of morphology (Kaplan and Kay, 1994; Beesley and Karttunen, 2003)[11].

Moreover, residual accuracies revealed that there is a division of labor among computational models with respect to linear and nested morphological structures. For instance, sequential models, such as Morpheme Markov Model and HMM accurately explained linear morphological structures at the expense of nested morphological structures. In other words, those sequential models were optimized to linear morphological structures, which naturally follows from their architecture where morphologically complex words are processed incrementally from left to right: linear morphological structures (e.g., *digit-al-ly*) can be predicted from morpheme bigrams of first-second morphemes (e.g., *digit-al*) and second-third morphemes (e.g., *al-ly*) both attested in the training data, while nested morphological structures (e.g., *un-predict-able*) cannot, because morpheme bigrams of first-second morphemes (e.g., **un-predict*) never appear in the training data. In contract, the hierarchical model is better balanced and fitted equally well to both linear and nested morphological structures, hence the greater deviance accuracy. Methodologically, this new evaluation metric remains to be adopted in the sentence processing literature to explore the division of labor among computational models for various syntactic constructions (Frank and Bod, 2011; Fossum and Levy, 2012).

Furthermore, remember that novel morphologically complex words were created as *potential* but not necessarily *actual* words (Halle, 1973; Bauer, 2014) with zero surface frequencies in the CELEX lexical database (Baayen et al., 1995) and semantic selectional restrictions not explicitly controlled. To the extent that those morphologically complex words are not stored in the mental lexicon, but rather computed online from component morphemes (cf. Hay, 2003), the fact that humans judged nested morphological structures as acceptable itself constitutes evidence in favor of abstract hierarchical structures.

Finally, we conclude from the results above that there is no fundamental distinction between syntax and morphology, as advocated by the framework of Distributed Morphology (Halle and Marantz, 1993). In formal language theory, given the naive intuition that actual words are stored in the

[11] As an anonymous reviewer correctly pointed out, this conclusion only applies to finite-state *acceptor* models of morphology, but crucially not finite-state *transducer* models of morphology (Kaplan and Kay, 1994; Beesley and Karttunen, 2003), because finite-state transducers can approximate context-free languages of finite length (cf. Langendoen, 1975), such as morphologically complex nested words tested in this paper.

finite lexicon, morphology has been claimed to be finite (in linguistic performance) with respect to weak generative capacity (i.e., string sets generated by the grammar; Langendoen, 1981; Heinz and Idsardi, 2011) and, correspondingly, computationally implemented as finite-state models (Kaplan and Kay, 1994; Beesley and Karttunen, 2003). However, as Carden (1983) correctly pointed out, switching emphasis to strong generative capacity as being only relevant for linguistic theory (i.e., structure sets generated by the grammar; Everaert et al., 2015; Fukui, 2015), morphology turned out to be infinite (in linguistic competence), as exemplified by recursion (e.g., *anti-missile missile*) and center-embedding (e.g., *undeundestabilizablizeable*)[12]. Relatedly, the apparent finite-stateness of morphology gave the impression that morphology is specially sensitive to linear order, but hierarchical structure plays an important role both in syntactic and morphological processing, especially when resolving long-distance dependencies, such as subject-verb agreement in syntax (e.g., *apples on the table are…* vs. **the table are…*) and prefix-suffix potentiation in morphology (e.g., *enjoyable*, **joyable*). Namely, morphological processing can be regarded as syntactic processing within words.

To recapitulate, going back to the original research question, the results of our psycholinguistic experimentation and computational modeling converged on the conclusion that human morphological competence should be characterized by abstract hierarchical structures, and cannot be reduced to surface linear strings. This conclusion clearly corroborates the internalist view that the grammar generates hierarchical structures (Sprouse et al., 2018), but does not deny probabilities traditionally associated with linear strings (Lau et al., 2016) on the assumption that probability distributions can be defined over hierarchical structures like PCFGs (Yang, 2008). Importantly for the debate between internalist vs. externalist positions, here we advocate the middle position on the spectrum between the extreme internalist ("only grammars, no probabilities") and extreme externalist ("only probabilities, no grammars") positions in favor of the eclectic view (Yang, 2004) that grammars (competence) categorically define grammaticality, while probabilities (performance) gradiently affect acceptability.

Nevertheless, there remain several issues with our psycholinguistic experiments and computational models. First, for psycholinguistic experiments, only morphologically complex words (i.e., grammatical words) were tested in this paper, but morphologically complex nonwords (i.e., ungrammatical words) must be developed and tested in order to make the results maximally comparable to the previous literature (Lau et al., 2016; Sprouse et al., 2018). Second, for computational models, Character and Syllable Markov Models were evaluated as instances of "amorphous" models in this paper, but state-of-the-art "amorphous" models, such as Naive Discriminative Learning (Baayen et al., 2011) and Recurrent Neural Network (Kirov and Cotterell, 2018) should be employed and evaluated against human acceptability judgments. Finally, acceptability judgment is known as an offline time-insensitive experimental measure, which only reflects the output of language processing including extra-linguistic factors like working memory and world knowledge (Sprouse, 2007). In order to complement this methodological limitation, novel morphologically complex words developed in this paper must be tested with online time-sensitive experimental measures, such as lexical decision (cf. Oseki et al., 2019).

5. CONCLUSION

In conclusion, we investigated whether human morphological competence should be characterized by abstract hierarchical structures internally generated by the grammar or reduced to surface linear strings externally attested in large corpora. Specifically, we performed a crowdsourced acceptability judgment experiment on morphologically complex words that are (i) unattested with zero surface frequencies and (ii) trimorphemic with linear and nested morphological structures. Then, five computational models of morphological competence were constructed and evaluated against human acceptability judgments via the acceptability measure called *syntactic log-odds ratio*: Character Markov Model (Character), Syllable Markov Model (Syllable), Morpheme Markov Model (Morpheme), Hidden Markov Model (HMM), and Probabilistic Context-Free Grammar (PCFG). Our psycholinguistic experimentation and computational modeling converged on the conclusion that "morphous" computational models with morpheme units outperformed "amorphous" computational models without morpheme units and, importantly, PCFG with hierarchical structures most accurately explained human acceptability judgments via several evaluation metrics, especially for morphologically complex words with nested morphological structures. Those results strongly suggest that PCFG with hierarchical structures is the most "human-like" computational model of morphological competence and, therefore, human morphological competence should be characterized by abstract hierarchical structures internally generated by the grammar.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by New York University's Institutional Review Board

[12]Interestingly, Carden (1983) provided the elaborate context for the example *undeundestabilizablizeable* in order to "assist our feeble performance to reach something closer to the power of the underlying competence" as follows: "At present, gentlemen, we live with an apparently stable balance of terror. But that balance may at any time be de-stabilized by our opponents. As the leaders of a peace-loving state, our objective must be an un-destabilize-able balance. But now, just as we have begun to un-destabilize=able-ize the situation, our opponents have bent all their efforts to de-un=destabilize=able-ize our precarious balance. In our current negotiations, it will not be enough to require an un-destabilize-able balance; we must aim to create an un-de=undestabilizable=ize-able balance."

(IRB). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YO and AM conceived and designed the project, and revised the manuscript together. YO created the stimuli, conducted the experiment, implemented the computational models, performed the statistical analyses, and prepared the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by JSPS KAKENHI Grant Numbers JP18H05589 and JP19H04990 (YO) and the NYU Abu Dhabi Institute Grant Number G1001 (AM).

ACKNOWLEDGMENTS

We would like to thank reviewers of *Frontiers in Psychology* and the members of the Neuroscience of Language Lab (NeLLab) at New York University.

REFERENCES

Ackerman, F., and Malouf, R. (2013). Morphological organization: the low entropy conjecture. *Language* 89, 429–464. doi: 10.1353/lan.2013.0054

Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Baayen, H., Milin, P., Durdevic, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychol. Rev*. 118, 438–481. doi: 10.1037/a0023851

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

Bar-Hillel, Y., and Shamir, E. (1960). Finite-state languages: formal representations and adequacy problems. *Bull. Res. Council Israel* 8F, 155–166.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. 67, 1–48. doi: 10.18637/jss.v067.i01

Bauer, L. (2014). Grammaticality, acceptability, possible words and large corpora. *Morphology* 24, 83–103. doi: 10.1007/s11525-014-9234-z

Bauer, L., Lieber, R., and Plag, I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Beesley, K., and Karttunen, L. (2003). *Finite State Morphology*. Chicago, IL: CSLI Publications, University of Chicago Press.

Berwick, R. (2018). "Revolutionary new ideas appear infrequently," in *Syntactic Structures after 60 Years*, eds N. Hornstein, H. Lasnik, P. Grosz-Patel, and C. Yang (Berlin: Mouton de Gruyter), 177–193. doi: 10.1515/9781501506925-181

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.

Carden, G. (1983). The non-finite = state-ness of the word formation component. *Linguist. Inq*. 14, 537–541.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Earley, J. (1970). An efficient context-free parsing algorithm. *Commun. Assoc. Comput. Mach*. 13, 94–102. doi: 10.1145/362007.362035

Embick, D. (2012). Roots and features (an acategorial postscript). *Theor. Linguist*. 38, 73–89. doi: 10.1515/tl-2012-0003

Erlewine, M. Y., and Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Nat. Lang. Linguist. Theory* 34, 481–495. doi: 10.1007/s11049-015-9305-9

Everaert, M., Huybregts, M., Chomsky, N., Berwick, R., and Bolhuis, J. (2015). Structures, not strings: linguistics as part of the cognitive sciences. *Trends Cogn. Sci*. 19, 729–743. doi: 10.1016/j.tics.2015.09.008

Fossum, V., and Levy, R. (2012). "Sequential vs. hierarchical syntactic models of human incremental sentence processing," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (Montreal, QC), 61–69.

Frank, S., and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychol. Sci*. 22, 829–834. doi: 10.1177/0956797611409589

Fukui, N. (2015). "A note on weak vs. strong generation in human language," in *50 Years Later: Reflections on Chomsky's Aspects*, eds A. Gallego and D. Ott (Cambridge, MA: MITWPL), 125–132.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguist. Inq*. 4, 3–16.

Halle, M., and Marantz, A. (1993). "Distributed morphology and the pieces of inflection," in *The View From Building 20, Essays in Linguistics in Honor of Sylvain Bromberger*, eds K. Hale and S. Keyser (Cambridge, MA: MIT Press), 111–176.

Hay, J. (2003). *Causes and Consequences of Word Structure*. New York, NY: Routledge.

Heinz, J., and Idsardi, W. (2011). Sentence and word complexity. *Science* 333, 295–297. doi: 10.1126/science.1210358

Kaplan, R., and Kay, M. (1994). Regular Models of Phonological Rule Systems. *Computational Linguistics*. 20, 331–378.

Kirov, C., and Cotterell, R. (2018). Recurrent neural networks in linguistic theory: revisiting Pinker and Prince (1988) and the past tense debate. *Trans. Assoc. Comput. Linguist*., 651–665. doi: 10.1162/tacl_a_00247

Langendoen, T. (1975). Finite-state parsing of phrase-structure languages and the status of readjustment Rulesin grammar. *Linguist. Inq*. 6, 533–554.

Langendoen, T. (1981). The generative capacity of word-formation components. *Linguist. Inq*. 12, 320–322.

Lau, J. H., Clark, A., and Lappin, S. (2016). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cogn. Sci*. 41, 1202–1241. doi: 10.1111/cogs.12414

Libben, G. (2003). "Morphological parsing and morphological structure," in *Reading Complex Words*, eds E. Assink and D. Sandra (New York, NY: Kluwer), 221–239. doi: 10.1007/978-1-4757-3720-2_10

Libben, G. (2006). "Getting at psychological reality: on- and off-line tasks in the investigation of hierarchical morphological structure," in *Phonology, Morphology, and the Empirical Imperative*, eds G. Wiebe, G. Libben, T. Priestly, R. Smyth, and S. Wang (Taipei: Crane), 349–369.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguist*. 4, 521–535. doi: 10.1162/tacl_a_00115

Marantz, A. (2013). No escape from morphemes in morphological processing. *Lang. Cogn. Process*. 28, 905–916. doi: 10.1080/01690965.2013.779385

Marelli, M., and Baroni, M. (2015). Affixation in semantic space: modeling morpheme meanings with compositional distributional semantics. *Psychol. Sci*. 122, 485–515. doi: 10.1037/a0039267

Oseki, Y., Yang, C., and Marantz, A. (2019). "Modeling hierarchical syntactic structures in morphological processing," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (Minneapolis, MN), 43–52. doi: 10.18653/v1/W19-2905

Ott, D. (2017). Strong generative capacity and the empirical base of linguistic theory. *Front. Psychol*. 8:1617. doi: 10.3389/fpsyg.2017.01617

Pauls, A., and Klein, D. (2012). "Large-scale syntactic language modeling with treelets," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Jeju Island), 959–968.

Pereira, F. (2000). Formal grammar and information theory: together again? *Philos. Trans. R. Soc. A* 358, 1239–1253. doi: 10.1098/rsta.2000.0583

Pinker, S., and Ullman, M. (2002). The past and future of the past tense. *Trends Cogn. Sci*. 6, 456–462. doi: 10.1016/S1364-6613(02)01990-3

Plag, I., and Baayen, H. (2009). Suffix ordering and morphological processing. *Language* 85, 109–152. doi: 10.1353/lan.0.0087

Rabinar, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Sennhauser, L., and Berwick, R. (2018). "Evaluating the ability of LSTMs to learn context-free grammars," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Brussels), 115–124. doi: 10.18653/v1/W18-5414

Sproat, R. (1992). *Morphology and Computation*. Cambridge, MA: MIT Press.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1, 123–134.

Sprouse, J., Indurkhya, S., Yankama, B., Fong, S., and Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *Linguist. Rev*. 35, 575–599. doi: 10.1515/tlr-2018-0005

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Comput. Linguist*. 21, 165–201.

Yang, C. (2004). Universal grammar, statistics or both? *Trends Cogn. Sci*. 8, 451–456. doi: 10.1016/j.tics.2004.08.006

Yang, C. (2008). The great number crunch. *J. Linguist*. 44, 205–228. doi: 10.1017/S0022226707004999

Zirkel, L. (2010). Prefix combinations in English: structural and processing factors. *Morphology* 20, 239–266. doi: 10.1007/s11525-010-9151-8

Conflict of Interest: 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership