A computational evaluation of feature distortion and cue weighting in sentence comprehension

Himanshu Yadav

Cumulative doctoral dissertation

Submitted to the Faculty of Human Sciences of the University of Potsdam
in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Cognitive Science

Year of submission: 2023

University of Potsdam
Faculty of Human Sciences

Supervisors:
Prof. Dr. Shravan Vasishth, Department of Linguistics, University of Potsdam
Dr. Garrett Smith, Department of Linguistics, University of Potsdam

Reviewers:
Prof. Dan Parker, Department of Linguistics, The Ohio State University
Prof. Dr. Shravan Vasishth, Department of Linguistics, University of Potsdam

Date of oral defense: March 21, 2023

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-58505
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-585055

Abstract

Successful sentence comprehension requires the comprehender to correctly figure out who did what to whom. For example, in the sentence *John kicked the ball*, the comprehender has to figure out who did the action of kicking and what was being kicked. This process of identifying and connecting the syntactically-related words in a sentence is called dependency completion. What are the cognitive constraints that determine dependency completion? A widely-accepted theory is *cue-based retrieval*. The theory maintains that dependency completion is driven by a content-addressable search for the co-dependents in memory. The cue-based retrieval explains a wide range of empirical data from several constructions including subject-verb agreement, subject-verb non-agreement, plausibility mismatch configurations, and negative polarity items.

However, there are two major empirical challenges to the theory: (i) Grammatical sentences' data from subject-verb number agreement dependencies, where the theory predicts a slowdown at the verb in sentences like *the key to the cabinet was rusty* compared to *the key to the cabinets was rusty*, but the data are inconsistent with this prediction; and, (ii) Data from antecedent-reflexive dependencies, where a facilitation in reading times is predicted at the reflexive in *the bodybuilder who worked with the trainers injured themselves* vs. *the bodybuilder who worked with the trainer injured themselves*, but the data do not show a facilitatory effect.

The work presented in this dissertation is dedicated to building a more general theory of dependency completion that can account for the above two datasets without losing the original empirical coverage of the cue-based retrieval assumption. In two journal articles, I present computational modeling work that addresses the above two empirical challenges.

To explain the grammatical sentences' data from subject-verb number agreement dependencies, I propose a new model that assumes that the cue-based retrieval operates on a probabilistically distorted representation of nouns in memory (Article I). This hybrid distortion-plus-retrieval model was compared against the existing candidate models using data from 17 studies on subject-verb number agreement in 4 languages. I find that the hybrid model outperforms the existing models of number agreement processing suggesting that the cue-based retrieval theory must incorporate a feature distortion assumption.

To account for the absence of facilitatory effect in antecedent-reflexive dependencies, I propose an individual difference model, which was built within the cue-based retrieval framework (Article II). The model assumes that individuals may differ in how strongly they weigh a syntactic cue over a number cue. The model was fitted to data from two studies on antecedent-reflexive dependencies, and the participant-level cue-weighting was estimated. We find that one-fourth of the participants, in both studies, weigh the syntactic cue higher than the number cue in processing reflexive dependencies and the remaining participants weigh the two cues equally. The result indicates that the absence of predicted facilitatory effect at the level of grouped data is driven by some, not all, participants who weigh syntactic cues higher than the number cue. More generally, the result demonstrates that the assumption of differential cue weighting is important for a theory of dependency completion processes. This differential cue weighting idea was independently supported

by a modeling study on subject-verb non-agreement dependencies (Article III).

Overall, the cue-based retrieval, which is a general theory of dependency completion, needs to incorporate two new assumptions: (i) the nouns stored in memory can undergo probabilistic feature distortion, and (ii) the linguistic cues used for retrieval can be weighted differentially. This is the cumulative result of the modeling work presented in this dissertation.

The dissertation makes an important theoretical contribution: Sentence comprehension in humans is driven by a mechanism that assumes cue-based retrieval, probabilistic feature distortion, and differential cue weighting. This insight is theoretically important because there is some independent support for these three assumptions in sentence processing and the broader memory literature. The modeling work presented here is also methodologically important because for the first time, it demonstrates (i) how the complex models of sentence processing can be evaluated using data from multiple studies simultaneously, without oversimplifying the models, and (ii) how the inferences drawn from the individual-level behavior can be used in theory development.

Acknowledgments

I am extremely grateful to my supervisor Shravan Vasishth for inspiring and empowering me to do all this work. This dissertation is a result of the vision and training that Shravan gave me. He always had time for me, helped me navigate through tough phases, gave me freedom in my research, and taught me so many basic skills for doing sciences. Shravan's scientific journey, his hard work, his resilience, and his commitment to his vision of research has inspired me, like many others, and will continue to do so. I am thankful that I had the opportunity to work with such a visionary and process-oriented mentor.

I am also grateful to my second supervisor Garrett Smith for his continuous support and encouragement. Garrett always kept me in a positive frame of mind and brought new perspectives to our discussions. He taught me to approach a research problem from many different angles.

All the members of our research group at Potsdam have, in one way or the other, contributed to this work. I thank Titus von der Malsburg, Sol Lago, Dario Paape, Kate Stone, Pia Schoknecht, Paula Lissón, Dorothea Pregla, Daniela Mertzen, Anna Laurinavichyute, and João Veríssimo. Their passion for this area of research and insightful discussions have inspired me throughout my Ph.D.

I want to express special gratitude to Samar Husain, my supervisor at the time of my Master's. When I first met Samar, I had zero skills, no research aptitude, and no real direction, but he turned me into a researcher. He put so much time and effort into my development, helped me in every circumstance, always treated me with respect, and taught me some important work ethics. A guide like Samar in the initial years of your academic career is a blessing for anyone.

The work presented in this dissertation would have not been possible without the raw data provided by Colin Phillips, Brian W. Dillon, Sol Lago, Matthew Tucker, Matthew Wagers, Nicole Patson, and Matthew Husband. My heartfelt thanks to them. I would also thank the audience of my talks at the CogSci and the MathPsych conference whose valuable feedback improved the quality of my work. I am thankful to Deutscher Akademischer Austauschdienst (DAAD) who funded my doctoral studies allowing me to freely pursue my research interests.

I am grateful to my parents who made all the sacrifices to ensure my education. I also thank my friends who always supported me whenever I had to take difficult decisions in life. The belief my friends and family showed in me has helped me in crucial situations.

Finally, I want to remember Late Sh. Biharilal guruji, my primary school teacher. Without his teaching and life lessons, I would not be half the person I am today. Even when he is not around, he inspires me to become a better human being.

Contents

Chapter 1

Introduction

Humans possess a remarkable ability to compute meaning out of a linear stream of sounds or signs that constitute a sentence. What are the cognitive processes that underlie this sentence comprehension skill in humans? From years of research, we have understood that comprehending a sentence requires the comprehender to correctly figure out who did what to whom. For example, to comprehend the sentence *John kicked the ball*, one must work out who did the action of kicking and what was being kicked. This process of identifying the linguistically related words in a sentence is called *dependency completion*. In the sentence *John kicked the ball*, the comprehender has to complete two noun-verb dependencies: the dependency between the subject noun *John* and the verb *kicked* and the dependency between the verb *kicked* and the direct object *the ball*.

Dependency completion processes have been extensively investigated using controlled experiments on several types of dependencies. One dependency type that has received considerable attention in the last three decades is the subject-verb number agreement dependency. This dependency is important for understanding the comprehension process because the reader sometimes fails to register the ungrammaticality when a hard morphosyntactic constraint on the subject-verb dependency is violated. In languages like English, the subject noun must agree in number with the verb: the reader's linguistic knowledge warrants that one must write *The key was rusty* and not *The key were rusty*. In spite of this number agreement constraint, the reader is sometimes misled into believing that the sentences like (1) are grammatical.

(1) * The key to the cabinets were rusty.

Such subject-verb agreement violations as in (1) — first noted by Mann (1982) — are produced surprisingly often by native speakers of a language. For instance, in an English treebank, 12% of the subject-verb dependencies violates the number agreement constraint.[1] Similarly, in Spanish, the number mismatch is observed in around 5% cases. Why the native speakers would produce the ungrammatical utterances like (1)?

The agreement violations were first investigated in sentence production studies. A consistent observation from these studies is that of *agreement attraction*: the ungrammatical sentences like (2a) are produced more often than sentences like (2b) (Bock and Miller, 1991).

(2) a. **Ungrammatical, plural distractor**
* The key to the cabinets were rusty.

[1] The conditional probabilities were computed from the Corpora from the Web (COW) (Schäfer, 2015, Schäfer and Bildhauer, 2012).

b. **Ungrammatical, singular distractor**
* The key to the cabinet were rusty.

While both (2a) and (2b) are ungrammatical due to number agreement violation, what makes the sentence (2a) better than (2b) during production? A widely-accepted explanation comes from the Marking and Morphing theory (Eberhard et al., 2005). The theory assumes that in sentences like (2a), the plural feature of the non-subject noun *the cabinets* spreads to the subject noun the key, which makes the subject noun more plural in (2a) compared to (2b). Consequently, the plural verb is produced more often in (2a) than (2b). In other words, the number match between the non-subject noun and the verb causes the observed agreement attraction effect in production.

Agreement attraction has been robustly found across different experimental paradigms; examples are acceptability judgments: (Hammerly et al., 2019, Häussler, 2009, Schlueter et al., 2018); forced-choice response: (Lago and Felser, 2018, Staub, 2009, 2010); and reading studies (Dillon et al., 2013, Lago et al., 2015, Tucker et al., 2015). For the work presented in this dissertation, I use the reading time data from sentence comprehension experiments because of two reasons. First, I am interested in evaluating the competing theories of sentence comprehension in humans. Second, the reading data from number agreement dependencies present two interesting empirical puzzles that remain unsolved by the existing theories.

In the reading studies (self-paced reading and eyetracking) on number agreement processing, it has consistently been observed that the verb is read faster in sentences like (2a) — where a non-subject noun matches the verb in number feature — compared to a baseline sentence (2b), where a non-subject noun does not match the verb in number (e.g., Avetisyan et al., 2020, Dillon et al., 2013, Jäger et al., 2020, Lago et al., 2021, 2015, Tucker et al., 2015, Wagers et al., 2009). This speedup in reading times at the auxiliary verb *were* in sentence (2a) vs. (2b) is taken as an online correlate of grammaticality illusion and has been referred to as *agreement attraction effect* in reading. Figure 1.1 (left) shows the agreement attraction effects observed in 17 published studies on subject-verb number agreement processing.

What is the underlying cognitive process that leads to this robust agreement attraction phenomenon in reading? Two broad classes of theories exist, the representation distortion-based theories and the cue-based retrieval theories. While these existing theories can account for the agreement attraction effects, they all fail to explain some key aspects of the observed reading time data on number agreement.

Representation distortion-based theories assume that the feature representation of the nouns stored in memory can probabilistically change or be lost over time. An influential proposal based on representation distortion is the *feature percolation theory* (e.g., Bock and Eberhard, 1993, Eberhard, 1997). The assumption is that, in some proportion of trials, the plural feature of the attractor noun *cabinets* in (2a) percolates up to the singular-marked subject noun and changes its representation. In trials where this percolation occurs, the subject is now plural-marked and matches the verb in number, making the sentence seem grammatical compared to the ungrammatical baseline sentence (2b).

In contrast to representation distortion theories, the *cue-based retrieval theory* (Lewis and Vasishth, 2005, Lewis et al., 2006) assumes that the dependency between the subject and the verb is resolved via a content-addressable search in memory (McElree, 2003, McElree et al., 2003). When the verb is encountered, a search is triggered in memory using the feature specifications of the verb's arguments (e.g., a noun phrase with [+subject], [+plural] features) to identify the target chunk for dependency completion. These feature specifications used for searching the co-dependent in memory are called retrieval cues. In a sentence like (2a), the attractor noun *cabinets* matches the [+plural] cue at the verb *were*; this cue-feature match leads to an occasional misretrieval of

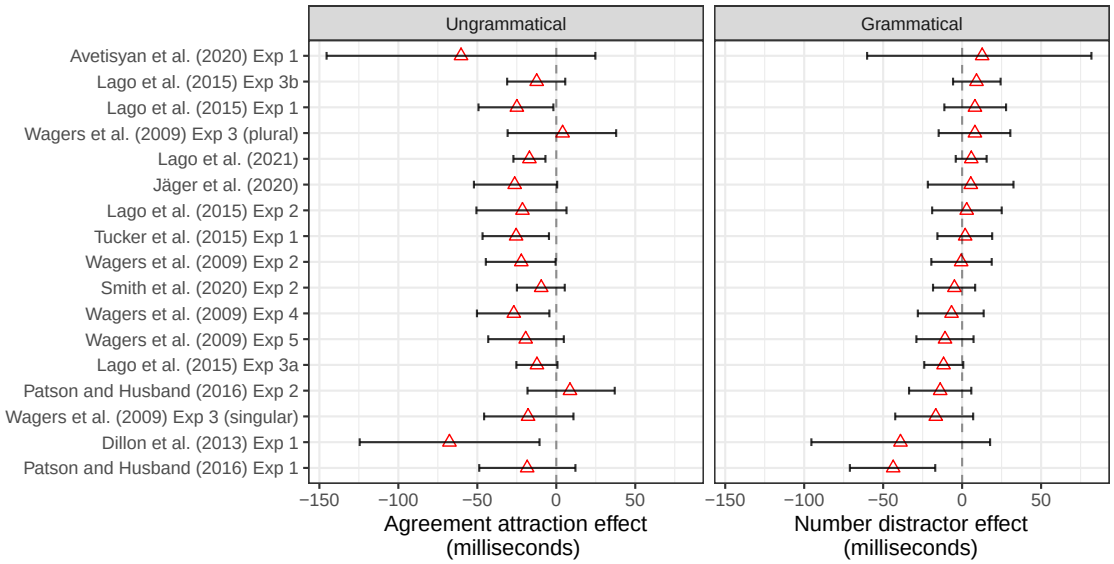


Figure 1.1: The pattern of effects in ungrammatical and grammatical subject-verb number agreement dependencies across 17 published datasets. The agreement attraction effect is the reading time (at the verb) in condition (2a) minus (2b); the number distractor effect is the reading time in (3a) minus (3b). The red triangles represent the estimated mean effects and the errors bars represent the 95% credible intervals.

the attractor noun (Patson and Husband, 2016, Wagers et al., 2009). This misretrieval is assumed to cause the illusion of grammaticality and faster reading times at the verb in (2a) vs. (2b) (see Engelmann et al., 2019).

As shown in the left panel of Figure 1.1, for ungrammatical sentences, the qualitative predictions of both the distortion-based and the retrieval-based theories are consistent with the data: across the 17 studies, the estimated agreement attraction effect — the difference in reading times between (2a) and (2b) — tends to be negative with some variation in the magnitude of the effect.

However, even though the existing theories can explain the ungrammatical sentences data from subject-verb number agreement dependencies, these theories are challenged by two important pieces of data:

1. The existing theories fail to explain the observed effects in the corresponding grammatical sentences shown in (3a,b). For example, in sentence (3a), the number distractor *cabinet* may either causes a slowdown, a speedup, or no effect at the verb compared to (3b) (see Figure 1.1; right panel). The existing theories — that can explain the attraction effect in ungrammatical sentences — cannot capture this *number distractor effect* in grammatical sentences. Thus, no existing theory of number agreement can explain the grammatical as well as ungrammatical sentences' data simultaneously.

2. Another kind of number agreement dependency —the antecedent-reflexive dependency— does not show an agreement attraction effect. For example, in the sentence *the bodybuilder who worked with the trainers injured themselves*, the cue-based retrieval theory predicts a speedup in reading times at the reflexive *themselves*, when compared with *The bodybuilder who worked with the trainer injured themselves*, but the data are inconsistent with this prediction.

My modeling work aims to address the above two empirical challenges. I discuss them in detail one by one.

1.1 The number agreement effect in grammatical sentences

While the existing theories of number agreement processing can account for the agreement attraction effects observed in ungrammatical sentences (see Figure 1.1; left panel), these theories falter when it comes to explaining the observed reading time pattern in the grammatical sentences shown in (3a,b). The estimated difference in reading time between (3a) and (3b) at the auxiliary verb fluctuates between −50 and 25 milliseconds across the 17 studies (see Figure 1.1), with some studies showing a slowdown in (3a) vs. (3b), and some studies a speed up. For convenience, I will refer to this difference in reading time between (3a) and (3b) as the *number distractor effect*.

(3) a. **Grammatical, singular distractor**
The key to the cabinet was rusty.

b. **Grammatical, plural distractor**
The key to the cabinets was rusty.

The number distractor effect—which is sometimes negative, sometimes positive, and often close to zero ms, depending on the study—cannot be fully captured by any of the existing theories.

The feature percolation theory —one of the representation distortion-based theories— maintains that in (3b), the plural feature on the distractor *cabinets* probabilistically percolate up to the subject *key*, resulting in the sentence being perceived as ungrammatical in a proportion of trials. Consequently, the average reading time at the verb in (3a) is predicted to be faster compared to (3b). This prediction is not entirely supported by the data: several studies show effectively no difference between (3a) and (3b), and some even suggest that (3a) is slower than (3b).

The cue-based retrieval theory makes the opposite qualitative prediction. Under the cue-based retrieval assumption, the auxiliary verb in (3a) searches for a noun with the [+subject] and [+singular] features but it takes longer to retrieve the subject noun due to the presence of a distractor noun *cabinet* with [+singular] features. Therefore, this partially-matching distractor is predicted to cause a slowdown at the verb (3a) compared to (3b), where no other noun partially matches the retrieval cues. Figure 1.1 shows that the predicted slowdown in (3a) vs. (3b) is not consistent with the estimates from the individual studies.

Thus, even though the predictions of both the representation distortion and cue-based retrieval are largely consistent with the agreement attraction effect in ungrammatical sentences, neither type of account can convincingly account for the observed range of number distractor effects in grammatical sentences. There seems to be no existing theory that can fully explain the qualitative pattern of number agreement effects.

Even though neither class of theories can fully account for the data, we can still investigate which model performs quantitatively better compared to the other candidate models. The 17 published datasets on number agreement allow us to evaluate competing models to understand how well the representation distortion theories and the cue-based retrieval theory perform when pitted against each other.

In Article I, we evaluate the relative fit of two retrieval-based models, three distortion-based models, and two hybrid models that combine representation distortion and the cue-based retrieval mechanism. We find that a new theory that integrates probabilistic feature percolation within the cue-based retrieval process shows the best predictive performance compared to all other

models considered. This finding suggests that subject-verb number agreement processing is driven by a cue-based retrieval process that operates on a probabilistically distorted representation of the nouns stored in memory. The main theoretical insight from Article I is that a general theory of dependency completion should incorporate the following two assumptions: (i) the co-dependents are identified and connected together via a content-addressable search in memory, and (ii) the representation of the co-dependents stored in memory gets probabilistically distorted over time.

1.2 The absence of agreement attraction in antecedent-reflexive dependencies

The second empirical challenge comes from a different kind of number agreement dependency: the antecedent-reflexive dependency. Consider the sentences (4a) and (4b). These sentences are ungrammatical because the dependency between the antecedent *bodybuilder* and the reflexive *themselves* violates the number agreement constraint. The existing theories of number agreement, e.g., the cue-based retrieval, predict the agreement attraction effect: the reading times at the reflexive are predicted to be faster in (4a) — where a non-antecedent noun matches the reflexive in number feature — compared to (4b), where none of the nouns matches the reflexive in number. However, the reading time data are inconsistent with this prediction.

(4) a. The bodybuilder who worked with the trainers injured themselves ...

b. The bodybuilder who worked with the trainer injured themselves ...

Dillon et al. (2013) demonstrated that the reflexive dependencies are immune to the agreement attraction effect. They attribute the absence of agreement attraction effect to Principle A of the binding theory (Chomsky, 1981), which states that an anaphor (e.g., a reflexive) must be bound within its governing category (e.g., its clause). Thus, in sentences (4a) and (4b), the antecedent would be a noun phrase that essentially c-commands the reflexive *themselves*. Following Sturt (2003), Dillon and colleagues argued that the search for an antecedent for the reflexive is guided exclusively by Principle A of the binding theory implying that the number marking on the reflexive *themselves* is not used as a retrieval cue for these dependencies. Several other researchers have taken a less extreme position; they hypothesized that the c-command cue dominates the number cue in processing antecedent-reflexive dependencies (Cunnings and Sturt, 2014, Kush, 2013, Parker and Phillips, 2017). I refer to these explanations collectively as the *cue weighting explanation*: the syntactic cue is weighting higher than the number cue in processing reflexive dependencies. An empirical challenge to the cue-weighting explanation comes from a large-scale replication of the Dillon et al. study by Jäger et al. (2020). Jäger and colleagues find that the reflexive dependencies do show a weak agreement attraction effect.

An important point missing from the above debate is that the individuals differ qualitatively in their effects in the case of reflexive dependencies: some participants do show an agreement attraction effect but some do not, and interestingly, this pattern persists across the two studies (see Figure 1.2). Given the differences in attraction effect estimates, it is possible that some, but not all, participants weigh the c-command cue higher than the number cue and the data from such participants might be deriving the absence of agreement attraction effect in the reflexives. This possibility raises an important question: Can individual differences in cue-weighting explain the observed pattern of attraction effects across the two studies on antecedent-reflexive dependencies?

Article II investigates this question. For the studies that provided data on reflexive dependencies (Dillon et al., 2013, Jäger et al., 2020), we found that only one-fourth of the participants

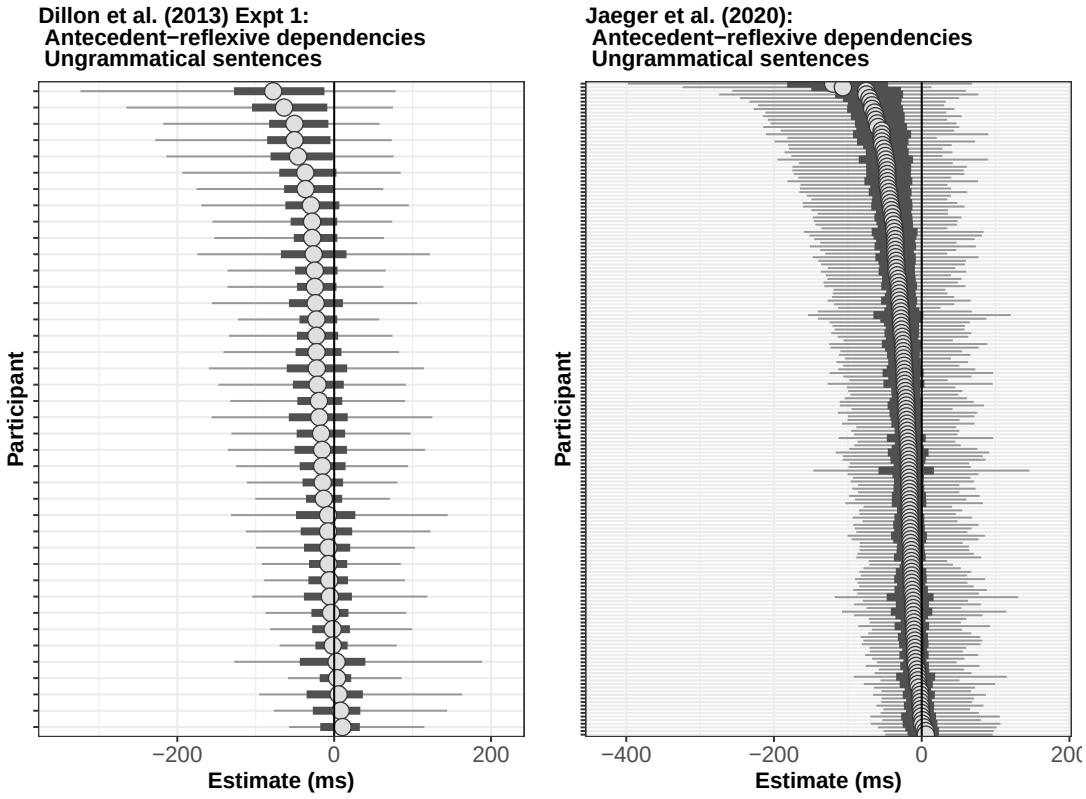


Figure 1.2: The individual-level agreement attraction effects observed for antecedent-reflexive dependencies in the two studies, Dillon et al. (2013) and Jäger et al. (2020). Under the cue-based retrieval account, the attraction effects should be negative. The individuals whose effect estimates are close to zero ms support the cue-weighting assumption: the structural cue is weighted higher than the number cue. The individuals with negative effect estimates (of smaller than −7 ms) support the assumption that the cues are weighted equally.

did have higher weighting for the c-command cue than the number cue in these dependencies, but also that most participants had equal weights for the two cues. The finding suggests that some participants do indeed strongly adhere to Principle A during online sentence comprehension, but the majority of participants weigh structural and non-structural cues approximately equally. Overall, the results suggest that the cue-weighting hypothesis only holds for a subset of English native speakers.

A broader conclusion from the modeling work in Article II is that the population-level effects inferred from a sample can mask theoretically important variation at the individual level. While population-level effects from Dillon et al. (2013) and Jäger et al. (2020) produced theoretically different conclusions, the individual-level effects reveal that the results from the two studies are consistent: Approximately one-fourth of the individuals in the population weigh syntactic cue higher than the number cue in processing reflexive dependencies.

Overall, Article II shows that the cue-weighting assumption — higher weighting of the syntactic cue over the non-syntactic cue — within the cue-based retrieval model could explain the apparent absence of agreement attraction effect in antecedent-reflexive dependencies. The importance of the cue-weighting assumption was further demonstrated in Article III, where we find that the higher weighting of the syntactic cue over the semantic cue can explain the processing of

subject-verb non-agreement dependencies in German vs. English.

The collective theoretical insight from my three articles is that a theory of dependency completion process should incorporate the following three assumptions: (i) dependency completion is driven by a content-addressable search in memory, (ii) the feature representation of the linguistic input stored in memory gets probabilistically distorted over time, and (iii) the linguistic cues used in dependency completion can be weighted differentially.

Chapter 2

List of articles

The following three articles constitute the core of this dissertation. All three articles contribute to understanding the cognitive processes that underlie dependency completion. The first article addresses the question of which theoretical assumptions can best explain the observed pattern of effects in subject-verb number agreement dependencies. The second article departs from the conventional approach of studying the average behavior and shows how individual difference modeling can be used to draw inferences for a theoretical question. The third article further demonstrates the utility of the individual difference approach in studying cross-linguistic processing differences.

Article I

Himanshu Yadav, Garrett Smith, Sebastian Reich, and Shravan Vasishth

Journal of Memory and Language, 129:104400, 2023

This article makes two main contributions to the study of dependency completion processes. First, the results show that a general theory of dependency completion should consider two crucial assumptions: (i) the co-dependents are identified and linked together via a content-addressable search in memory, and (ii) the feature representation of the co-dependents gets probabilistically distorted when they are stored in memory. The second contribution is methodological. The article presents a Bayesian method to evaluate complex models of sentence processing using data from multiple studies simultaneously, without compromising the complexity of the models. The algorithms developed here can be easily adapted to compare computational models of any underlying cognitive process.

Article II

Himanshu Yadav, Dario Paape, Garrett Smith, Brian Dillon, and Shravan Vasishth

Open Mind, 6:147–168, 2022

This article reveals an important theoretical insight that the linguistic cues used for dependency completion can be weighted differentially by a comprehender. We find that some comprehenders weigh the syntactic cue higher than the number cue, while some others weigh the two cues equally in processing antecedent-reflexive dependencies. The result suggests that the individuals

differ in how they weigh certain linguistic cues over the other cues. Moreover, the article demonstrates the theoretical significance of modeling individual differences. In two previous studies that investigated reflexive dependencies, the average behavior lead to theoretically different interpretations. But the distribution of individual-level behavior was consistent across the two studies: Only one-fourth of the participants in both studies weighed the syntactic cue higher than the number cue.

Article III

Himanshu Yadav, Garrett Smith, Daniela Mertzen, Ralf Engbert, and Shravan Vasishth

Proceedings of the Annual Meeting of the Cognitive Science Society, 44, 2022

This short article provides an independent illustration of how individual differences can be used for drawing theoretical inferences. The main finding is that English and German participants differ in how they weigh syntactic cues relative to semantic cues. While most English participants weigh the syntactic cue and the animacy cue equally, a majority of German participants weigh the syntactic cue higher. The article supports the broader hypothesis that the native speakers of a particular language may learn to use certain cues more strongly and reliably over the others which may lead to cross-linguistic differences in processing.

Chapter 3

Article I

Number feature distortion modulates cue-based retrieval in reading

Himanshu Yadav, Garrett Smith, Sebastian Reich, and Shravan Vasishth
Journal of Memory and Language, 129:104400, 2023

DOI: https://doi.org/10.1016/j.jml.2022.104400
Code: https://osf.io/gqj3p/

Journal of Memory and Language 129 (2023) 104400

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Number feature distortion modulates cue-based retrieval in reading

Himanshu Yadav [a,*], Garrett Smith [a], Sebastian Reich [b], Shravan Vasishth [a]

[a] Department of Linguistics, University of Potsdam, Germany
[b] Institute for Mathematics, University of Potsdam, Germany

ARTICLE INFO

Keywords:
Agreement attraction
Encoding interference
Cue-based retrieval
Lossy memory representations
Feature percolation
Subject–verb number agreement

ABSTRACT

In sentence comprehension, what are the cognitive constraints that determine number agreement computation? Two broad classes of theoretical proposals are: (i) *Representation distortion accounts*, which assume that the number feature on the subject noun gets overwritten probabilistically by the number feature on a non-subject noun, leading to a non-veridical memory trace of the subject noun; and (ii) The *cue-based retrieval account*, a general account of dependency completion processes which assumes that the features on the subject noun remain intact, and that processing difficulty is only a function of the memory constraints on dependency completion. However, both these classes of model fail to account for the full spectrum of number agreement patterns observed in published studies. Using 17 benchmark datasets on number agreement from four languages, we implement seven computational models: three variants of representation distortion, two cue-based retrieval models, and two hybrid models that assume both representation-distortion and retrieval. Quantitative model comparison shows that the best fit is achieved by a hybrid model that assumes both feature distortion (specifically, feature percolation) and cue-based retrieval; numerically, the second-best quantitative fit was achieved by a distortion-based model of number attraction that assumes grammaticality bias during reading. More broadly, the work furnishes comprehensive evidence to support the idea that cue-based retrieval theory, which aims to be a general account of dependency completion, needs to incorporate a feature distortion process.

Introduction

Successful sentence comprehension requires that the reader correctly work out who did what to whom. To do so, the reader needs to identify the syntactic relations between words, a process called dependency completion. Dependency completion is a key step in sentence processing because a prerequisite for comprehending a sentence is that dependencies between nouns and verbs are resolved. The subject–verb dependency is especially interesting because, surprisingly, a hard morphosyntactic constraint on this dependency can sometimes be violated without the comprehender even registering the resulting ungrammaticality: In languages like English, the number marking on the subject must agree in number with the verb: the reader's internalized grammar stipulates that one must write *The key was rusty* and not *The key were rusty*. In spite of this, the reader is sometimes misled into thinking that the following sentence is grammatical: *The key to the cabinets were rusty*. Such subject–verb agreement mismatches – colorfully labeled atmosphere effects (Mann, 1982) – are produced surprisingly often by native speakers.

Agreement attraction was first observed in sentence production studies: participants produce sentences like (1a) more frequently than (1b) (Bock & Miller, 1991). Since then, the effect has been robustly found across different experimental paradigms; examples are acceptability judgments: (Hammerly, Staub, & Dillon, 2019; Häussler, 2009; Schlueter, Williams, & Lau, 2018); forced-choice response: (Lago & Felser, 2018; Staub, 2009, 2010b); and reading (self-paced reading and eyetracking) studies (Dillon, Mishler, Sloggett, & Phillips, 2013; Lago, Shalom, Sigman, Lau, & Phillips, 2015). In this paper, we focus on reading time data from sentence comprehension experiments because we are interested in evaluating the predictions of competing sentence comprehension models. In these reading studies (self-paced reading and eyetracking) using several languages, it has consistently been shown that the verb is read faster in sentences like (1a) – where a non-subject noun matches the verb in number feature – compared to an equally ungrammatical baseline sentence (1b), where a non-subject noun does not match the verb in number (e.g., Avetisyan, Lago, & Vasishth, 2020; Dillon et al., 2013; Jäger, Mertzen, Van Dyke, & Vasishth, 2020; Lago, Acuña Fariña, & Meseguer, 2021; Lago, et al., 2015; Tucker, Idrissi, & Almeida, 2015; Wagers, Lau, & Phillips, 2009). The faster reading time observed at the auxiliary verb *were* in sentence (1a) is usually taken as an online correlate of a failure to register the ungrammaticality,

* Corresponding author.
E-mail address: hyadav@uni-potsdam.de (H. Yadav).

https://doi.org/10.1016/j.jml.2022.104400
Received 4 April 2022; Received in revised form 2 December 2022; Accepted 6 December 2022

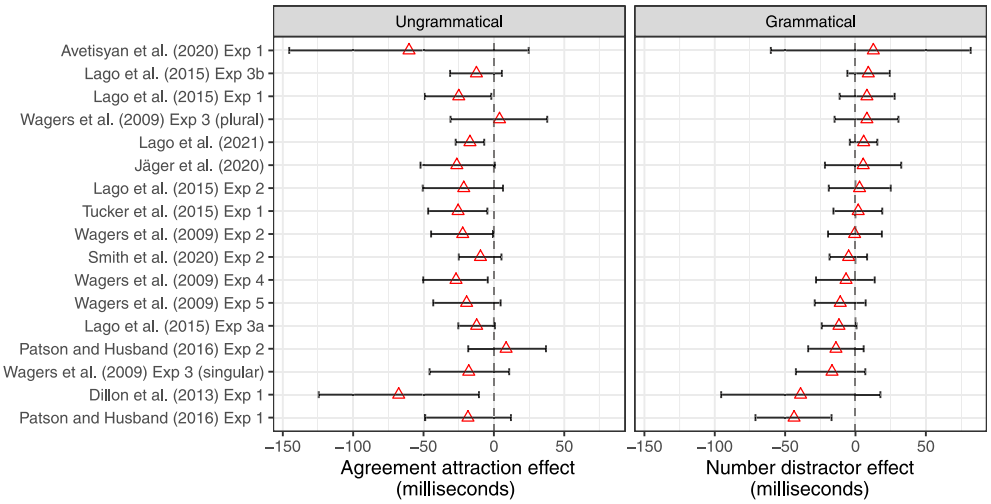


Fig. 1. The pattern of effects in ungrammatical and grammatical subject–verb number agreement dependencies across 17 published datasets. The agreement attraction effect is the reading time (at the verb) in condition (1a) minus (1b); the number distractor effect is the reading time in (2a) minus (2b). The red triangles represent the estimated mean effects and the errors bars represent the 95% credible intervals.

i.e., of an illusion of grammaticality. The phenomenon has in recent years come to be referred to as *agreement attraction*. Fig. 1 (left) shows the number agreement effects observed in 17 studies.

1. (a) **Ungrammatical, plural distractor**
   * The key to the cabinets were rusty.
   (b) **Ungrammatical, singular distractor**
   * The key to the cabinet were rusty.

What is the underlying cognitive process that leads to this remarkable illusion of grammaticality? Several theories exist, but all of them fail to explain some key aspect or the other of the published reading-time data on number agreement.

Two broad classes of theory offer an account for the agreement attraction illusion. *Representation distortion accounts* assume that the number-feature representation of the subject noun stored in memory can probabilistically change or even be lost as a function of time. An influential representation distortion account is the *feature percolation theory* (e.g., Bock & Eberhard, 1993; Eberhard, 1997). The claim is that, in some proportion of trials, the plural feature of the distractor noun *cabinets* in (1a) percolates up to and overwrites the singular-marked subject noun phrase. In trials where this percolation occurs, the subject is now plural-marked and matches the verb in number, making the sentence seem grammatical compared to the ungrammatical baseline sentence. One further representation distortion account that we will consider later is the recently proposed lossy compression model (Futrell, Gibson, & Levy, 2020), which takes into consideration how likely different verb number markings are given a noisy memory representation of the preceding words. This model has never been used to account for agreement attraction, but it can explain aspects of number agreement processing. The predictions of lossy compression for agreement phenomena will be discussed in the modeling section below.

In contrast to representation distortion accounts, the *cue-based retrieval theory* (Lewis & Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006) assumes that subject–verb dependency completion is carried out via a content-addressable search in memory (McElree, 2003; McElree, Foraker, & Dyer, 2003). This search is triggered when the verb is processed; the verb relies on the feature specifications of its arguments to search for the relevant phrase (e.g., a noun phrase) in memory. The retrieval triggered at the verb uses feature specifications such as [+subject] and [+plural] to seek out the relevant syntactic argument. These feature specifications used for searching memory are called retrieval cues. In a sentence like (1a), the distractor noun phrase *cabinets* matches the [+plural] cue at the verb *are*; this match occasionally leads to a misretrieval of the distractor noun (Patson & Husband, 2016; Wagers et al., 2009). This misretrieval is assumed to cause the illusion of grammaticality and faster reading time (for the technical details of this misretrieval process, see Engelmann, Jäger, & Vasishth, 2019).

As shown in the left panel of Fig. 1, for ungrammatical sentences, the qualitative predictions of both these types of model seem to be correct: across the 17 studies,[1] the estimated agreement attraction effect — the difference in reading time between (1a) and (1b) — tends to be negative with some variation in the magnitude of the effect.[2]

However, even though both classes of theory can explain the ungrammatical sentences discussed above, all existing theories falter when it comes to explaining the observed reading time pattern in the grammatical sentences shown in (2a,b). The estimated difference in reading time between (2a) and (2b) at the auxiliary verb fluctuates around zero ms across the 17 studies (see Fig. 1), with some studies showing a slowdown in (2a) vs. (2b), and some studies a speed up. For convenience, we will refer to this fluctuating difference in reading time between (2a) and (2b) as the *number distractor effect*, to distinguish it from the agreement attraction effect (we will refer to the two classes of effect – agreement attraction and the number distractor effect – collectively as *number agreement effects*).

2. (a) **Grammatical, singular distractor**
   The key to the cabinet was rusty.
   (b) **Grammatical, plural distractor**
   The key to the cabinets was rusty.

[1] These 17 studies were chosen because these are the only ones for which we have the original data; having the data made it possible to compute the estimates and their uncertainty in a unified manner using maximally conservative Bayesian linear mixed models (Schad, Betancourt, & Vasishth, 2021).

[2] We focus here on modeling the observed effects and their uncertainty, instead of focusing on whether the individual results were statistically significant or not. See Vasishth and Gelman (2021) and Kruschke and Liddell (2018) for the motivation for this approach.

The number distractor effect – which is sometimes negative, sometimes positive, and often close to zero ms, depending on the study – cannot be fully captured by any of the existing accounts.[3]

The feature percolation account maintains that in (2b), the plural feature on the distractor *cabinets* should probabilistically percolate up to the subject *key*, resulting in the sentence being perceived as ungrammatical. As a result, the reading time at the auxiliary verb in (2a) is predicted to be faster compared to (2b); the sign of the effect at the auxiliary verb (2a minus 2b) should be negative. This prediction is not entirely supported by the data: several of the individual studies show effectively no difference between (2a) and (2b), and some even suggest that (2a) is slower than (2b).

The cue-based retrieval theory makes the opposite qualitative prediction to the feature percolation model. This prediction is also inconsistent with the data. In this account, the auxiliary in (2a) should be read slower than in (2b). The slowdown at the verb is predicted to occur because in (2a) the verb searches for a noun with the [+subject] and [+singular] features but takes longer to retrieve the subject due to the fan effect (Anderson, et al., 2004; Schneider & Anderson, 2012). Under this view, the sign of the effect (2a minus 2b) should be positive. Fig. 1 shows that the prediction of a positive sign on the effect is not consistent with the estimates from the individual studies.

Thus, even though the predictions of both the representation distortion account (here, for expository purposes, we are only focusing on feature percolation) and cue-based retrieval are largely consistent with the observed agreement attraction effect, neither type of account can convincingly account for the number distractor effects in these 17 studies. This leaves us at an impasse: there seems to be no model that can explain the full pattern of effects from number agreement studies.[4]

Even though neither class of model can fully account for the data, it is still informative to ask which model performs better compared to the competitor models. The 17 datasets give us a unique opportunity to carry out a model comparison to understand how well the representation distortion models and the cue-based retrieval models perform when pitted against each other. For model comparison, it is critical to evaluate models via their relative predictive fit on unseen data. Evaluating model fit on unseen data is important because the alternative – evaluating model fit using the same data that the model was trained on – will lead to overfitting.

Given these considerations, the approach we will take is as follows: each of the models' numerical parameters will be fit to the estimates from a subset of the 17 studies taken together, holding out the remaining studies' estimates. Then, the quality of predictive fit to the held-out estimates will be used to quantify relative model fit. This approach – widely used in machine learning under the rubric of cross-validation (Vehtari, Gelman, & Gabry, 2017; Vehtari, Ojanen, et al., 2012) – yields a measure of the relative predictive accuracy of the competing models. Cross-validation is especially useful when comparing very different models.

Using this approach, we evaluate the relative fit of two retrieval-based models, three distortion-based models, and two hybrid models that combine representation distortion and the cue-based retrieval mechanism. Anticipating our main result, the model comparison leads us to a new theory that embeds feature percolation within the cue-based retrieval process. This new theory, which shows the best predictive accuracy compared to all other models considered, assumes that cue-based retrieval at the verb operates on a probabilistically distorted representation of the pre-verbal input.

[3] The non-linear cue-combination proposals in Wagers et al. (2009) and Parker (2019) can explain the close-to-zero effect in grammatical sentences. However, they fail to fully capture the agreement attraction effect in ungrammatical sentences; we illustrate this prediction in the modeling section.

[4] However, a recent transformer-based model by Ryu and Lewis (2021) has been shown to capture the typical number agreement effects in English for both grammatical and ungrammatical sentences; we have not considered this model here because this would take us far beyond the scope of the present paper. But it would be a good candidate model for future work focusing on English data.

Seven models of number agreement attraction

We first discuss the assumptions and predictions of the existing retrieval-based and distortion-based models, and then we present our hybrid model proposals.

For each model, we use a Bayesian approach to generate predictions for the number agreement effects. The Bayesian method allows us to compute a distribution of reading times and of the effects from a model based on our prior assumptions about the parameters of the model. First, a prior distribution is defined on the free parameters of the model; this prior distribution reflects our knowledge, beliefs, or assumptions about the plausible values of the parameter. Second, parameter values are repeatedly drawn from the prior distribution and used to generate reading times, from which the predicted effects (the agreement attraction effect, and the number distractor effect) can be derived. These simulations produce a distribution of predicted reading times called the *prior predictive distribution*. In this section, we compare the prior predictive distributions of each model with the estimates for the agreement attraction and number distractor effects. Then, in the following section, we test the models' predictive accuracy on held-out data.

Cue-based retrieval models

The cue-based retrieval models assume that the dependency completion between the subject and the verb is driven by a content-addressable search in memory using *retrieval cues* at the verb. For example, to identify the correct subject noun in sentences like *the key to the cabinets was rusty*, the verb could use retrieval cues like [subject], [singular] etc. The retrieval-based models may differ in how these individual cues combine to search for the target noun. Some models assume that the cues are combined linearly such that each cue contributes independently in the retrieval process (Anderson, et al., 2004; Anderson & Lebiere, 2014). By contrast, some models assume the non-linear cue-combination: the contribution of a retrieval cue is not independent of the other cues (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981). We first present the cue-based retrieval model of Lewis and Vasishth (2005) which assumes that the cues are combined linearly.

The cue-based retrieval model of Lewis and Vasishth (2005)

The cue-based retrieval theory described in Lewis and Vasishth (2005) is computationally implemented and has been extensively tested on a variety of constructions (Engelmann et al., 2019; Vasishth & Engelmann, 2022; Vasishth, Nicenboim, Engelmann, & Burchert, 2019). In this paper, we consider the cue-based retrieval model as implemented in Engelmann et al. (2019).

The cue-based retrieval model assumes that

(1) Dependency completion between the subject and the verb is driven by a content-addressable search of nouns in memory. Verbs carry feature specifications like [subject] or [plural]. These *retrieval cues* specify the features that nouns in memory should have in order to fill the role of the verb's subject.
(2) Each noun phrase that matches a retrieval cue receives a certain amount of activation, and the chunk with the highest activation gets retrieved for dependency completion.

(a) The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{+singular}$ $was^{subject}_{singular}$ rusty

(b) The $key^{+subject}_{+singular}$ to the $cabinets^{-subject}_{-singular}$ $was^{subject}_{singular}$ rusty

(c) * The $key^{+subject}_{-plural}$ to the $cabinets^{-subject}_{+plural}$ $were^{subject}_{plural}$ rusty

(d) * The $key^{+subject}_{-plural}$ to the $cabinet^{-subject}_{-plural}$ $were^{subject}_{plural}$ rusty

Fig. 2. Feature specifications at the nouns and the verb used by cue-based retrieval process. Each noun phrase receives activation proportional to the number of features it has that match the verb's retrieval cues. Features that match the retrieval cues are printed in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Consider the ungrammatical conditions (c) and (d) in Fig. 2. The verb *were* requires a noun phrase with [+plural] and [+subject] features to complete the subject–verb dependency. In condition (c), each noun phrase partially matches the retrieval cues at the verb, i.e., *the key* matches [subject] cue and *the cabinets* matches the [plural] cue. Consequently, a race for retrieval is initiated such that either noun phrase can be retrieved in each trial. This causes statistical facilitation in retrieval times (Raab, 1962) in condition (c) compared to condition (d), where only one noun phrase, *the key*, partially matches the retrieval cues. Thus, the model predicts a facilitatory effect in ungrammatical conditions, i.e., a speed-up in condition (c) compared to (d).

In the grammatical conditions shown in Fig. 2, the model predicts an inhibitory effect —a slowdown in condition (a) compared to (b). This slowdown is predicted to occur because of the *fan effect* (Anderson, et al., 2004; Schneider & Anderson, 2012): the singular distractor noun in (a), *the cabinet*, reduces the total amount of activation to be received by the singular subject noun compared to that in condition (b).

The details of the model are shown in section The cue-based retrieval model. We generate prior predictions from the model assuming, as is commonly done (Anderson, et al., 2004; Vasishth, 2020), one free parameter, the scaling parameter F. This parameter controls the overall speed of processing. This parameter has to be estimated from the observed data in order to map the model output onto the same scale as the human reading times. For the prior distribution on this scaling parameter F, we choose a truncated normal distribution

$$F \sim Normal_{lb=0.05}(0.15, 0.05) \quad (1)$$

where $lb = 0.05$ indicates a lower bound of 0.05 on scaling values. We choose this lower bound because the model generates unrealistically fast reading times when $F < 0.05$; $F > 0.05$ generates reading times that are more consistent with human reading data (see chapter 6, Nicenboim, Schad, & Vasishth, 2022). Section Choice of priors on the scaling parameter has a detailed discussion about how the priors on the scaling parameter were chosen. The prior predictions of the model given the above prior are shown in Fig. 3.

The non-linear cue-based retrieval model

The above cue-based retrieval model by Lewis and Vasishth (2005) assumes that the retrieval cues are combined linearly: the total activation received by a memory chunk is the sum of the activation spread via each individual cue. Under this assumption, as the number of matching cues increases, the total activation of the target chunk increases linearly. By contrast, some memory models assume a *non-linear cue combination* such that the total activation received by a chunk is the product of activation spread via each retrieval cue (Gillund & Shiffrin, 1984; Parker, 2019; Raaijmakers & Shiffrin, 1981; Van Dyke, 2007; Wagers, 2008). This multiplicative cue-combination implies that the contribution of an individual cue is not independent of the other cues. Wagers et al. (2009) argue that a cue-based retrieval model assuming non-linear cue combination can capture the number agreement effects in both grammatical and ungrammatical sentences. More specifically, they suggest that a direct access mechanism (McElree, 2000) with non-linear cue-combination can account for the typical pattern of close-to-zero effects in grammatical sentences. The idea is that the distractor noun in condition (a) (see Fig. 2) – which matches the number cue but not the subject cue – would cause almost zero interference in processing. This is because the activation spread by the [subject] cue to the distractor tends to be zero and therefore, the total activation received by the distractor via [subject] cue and the [number] cue would tend to zero if cues are combined non-linearly. Consequently, the distractor in grammatical sentences would have no effect on retrieval and the model would predict no difference between conditions (a) and (b) (Fig. 2).

Here, we implement a non-linear cue-based retrieval model (see Parker, 2019; Wagers, 2008) assuming direct access from memory (McElree, 2000). The assumptions of the model are as follows,

1. The dependency completion between the subject and the verb is driven by a content-addressable search for the subject noun.
2. Each noun in memory has a certain probability of retrieval determined by its degree of match with the retrieval cues.
   (a) The retrieval probability of a noun i is a non-linear function of its degree of match with the retrieval cues, i.e. $P_i = \frac{\prod_{j=1}^{N} s_{ij}^{W_j}}{\sum_{i=1}^{n} \prod_{j=1}^{N} s_{ij}^{W_j}}$. Where j indexes retrieval cues, i indexes nouns in memory, S_{ij} indicate the degree of match between a cue j and noun i (S_{ij} takes the value 0.99 if the noun matches the cue and the value 0.01 when it does not), W_j is the relative weight of cue j (W_j is set to 1 for all cues in our implementation).
   (b) The cost of retrieval is the same regardless of what was retrieved.
3. If retrieval fails, i.e., when the retrieved noun does not fully match the retrieval cues, then backtracking occurs with some probability such that the incorrectly retrieved noun is replaced by the target noun in a proportion of trials (Martin & McElree, 2008; Nicenboim & Vasishth, 2018).

However, it is not clear from the literature how to operationalize the backtracking assumption (Assumption (c)) for ungrammatical sentences. This is because in ungrammatical sentences, none of the nouns fully match the retrieval cues. The previous implementations of the direct access model for reading times have not specified what should happen in the ungrammatical sentences (see Lissón, et al., 2021; Nicenboim & Vasishth, 2018; Vasishth et al., 2019). The question is whether backtracking can occur for both the subject and the attractor noun, or whether it can occur only for the attractor noun.

Here, we assume that the backtracking can occur probabilistically if any noun does not fully match the retrieval cues. That is, in grammatical sentences, only the distractor can cause backtracking but in ungrammatical sentences, both the subject and the attractor can trigger backtracking.[5]

[5] We also explored the assumption that backtracking occurs only for the non-subject noun, but such a model makes highly inconsistent predictions with respect to observed effects in ungrammatical sentences (see section The non-linear cue-based retrieval model).

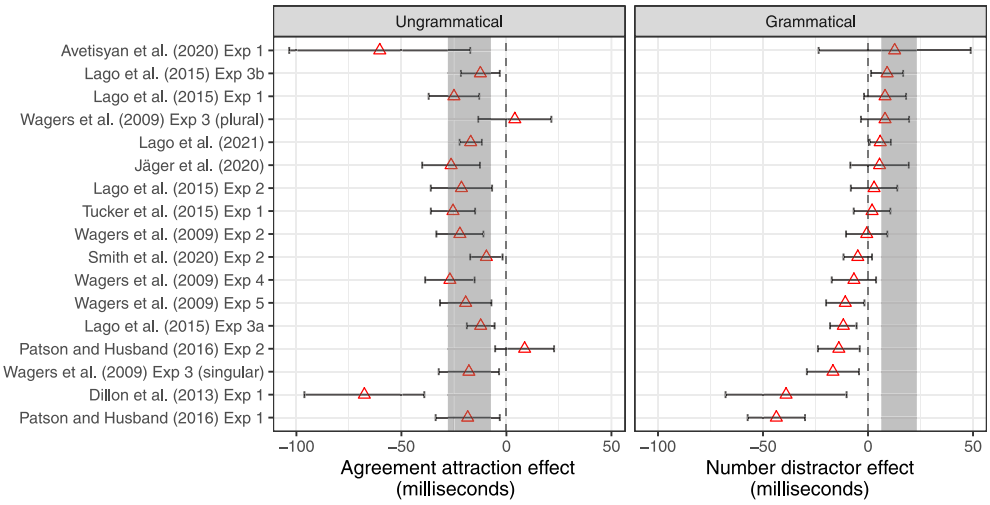


Fig. 3. Prior predictions of the cue-based retrieval model: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model. The red triangles and the error bars around them show the observed effects for each dataset specified on the *y*-axis.

Similar to the cue-based retrieval model with linear cues, we assume only one free parameter in the model, i.e., the scaling parameter S. We define the following prior on the scaling:

$$S \sim Normal_{lb=5.2}(5.3, 0.05) \quad (2)$$

This prior is a truncated normal distribution with mean 5.3, standard deviation 0.05, and a lower bound at 5.2. Similar to the cue-based retrieval model, we choose this lower bound on scaling because the model generates reasonable reading times under this constraint (see section Choice of priors on the scaling parameter).

The prior predictions of the model given the above priors are shown in Fig. 4. As predicted, the model correctly captures the close-to-zero number distractor effect in grammatical sentences but it fails to capture the agreement attraction effect (in ungrammatical sentences). In the ungrammatical sentence (c) (Fig. 2), the retrieval probability is 50%–50% for the subject and the attractor, either of them can be retrieved, but in condition (d), the subject noun has almost 100% retrieval probability. But backtracking can occur regardless of which noun is retrieved initially. Consequently, in both conditions (c) and (d), reading times are sampled from the same mixture of distributions. Thus, the model predicts no difference between conditions (c) and (d) with some uncertainty (see Fig. 4).

The variants of the non-linear cue-based retrieval model with slightly different assumptions, e.g., the model specified in Parker (2019), were also implemented (see section The non-linear cue-based retrieval model of Parker (2019)).[6] These variants also fail to capture the observed pattern in both grammatical and ungrammatical sentences; they make either similar or more inconsistent predictions compared with the above direct access-based model. See sections The non-linear cue-based retrieval model, The non-linear cue-based retrieval model of Parker (2019) for the implementational details and prior predictions of all non-linear cue-based retrieval models.

One of the possible assumptions that we have not explored here is that backtracking can occur only when the initially retrieved noun mismatches the verb in number feature. That is, it does not matter whether the initially retrieved noun is the subject or a non-subject; backtracking is triggered only when the noun's number feature is different from the verb's. A non-linear cue-combination model with such an assumption can potentially explain the agreement attraction pattern in ungrammatical sentences which cannot be fully explained by the other variants of the non-linear cue-based retrieval model. We have not tested such a model in this paper because such a model would require an additional assumption to explain the grammatical sentences' data that would be different from the assumption for ungrammatical sentences: backtracking in grammatical sentences can occur only when the retrieved noun mismatches in structural features with the verb. Thus, the model would have to make different assumptions for grammatical and ungrammatical sentences regarding what triggers the backtracking process. Such a model would go against our general constraint on the models implemented here: a single set of assumptions should underlie the number agreement effects in both grammatical and ungrammatical sentences.

Representation distortion-based models

Several models share the *representation distortion assumption*, which is that the representation of the linguistic input stored in memory can get corrupted or be lost with time. We call these models representation distortion-based models; the models differ in the assumed process by which distortion occurs. We implement three of these models, the feature percolation model, the marking and morphing model assuming grammaticality bias, and the lossy compression model.

All three representation distortion models that we implement (as well as the hybrid models discussed below) have a common free parameter that determines the rate of change in the representation of the subject and/or distractor nouns. We call this rate-of-change parameter the *distortion rate* parameter. The distortion rate determines the probability that a pre-verbal noun changes its number representation in a given trial. The way that this is implemented differs depending on the model (see below for details). In general, a distortion rate of 0.2 would mean that the representation changes in 20% of the trials. What would be a reasonable prior for the distortion rate parameter?

Agreement attraction errors observed in acceptability judgment studies (e.g., Hammerly et al., 2019; Häussler, 2009; Schlueter et al., 2018; Tanner, Nicol, & Brehm, 2014) can serve as an empirical basis for deciding on plausible prior values for the distortion rate parameter. In these studies, participants are given sentences like *(a) the key to cabinets were …* and *(b) the key to cabinet were …*, and they are asked to judge whether the given sentence is acceptable or not. A consistent observation is that the participants occasionally rate sentences like (a) and (b) as grammatically acceptable. But interestingly,

[6] These models adopt the retrieval time equation defined in the ACT-R architecture (Anderson, et al., 2004; Anderson & Lebiere, 2014).

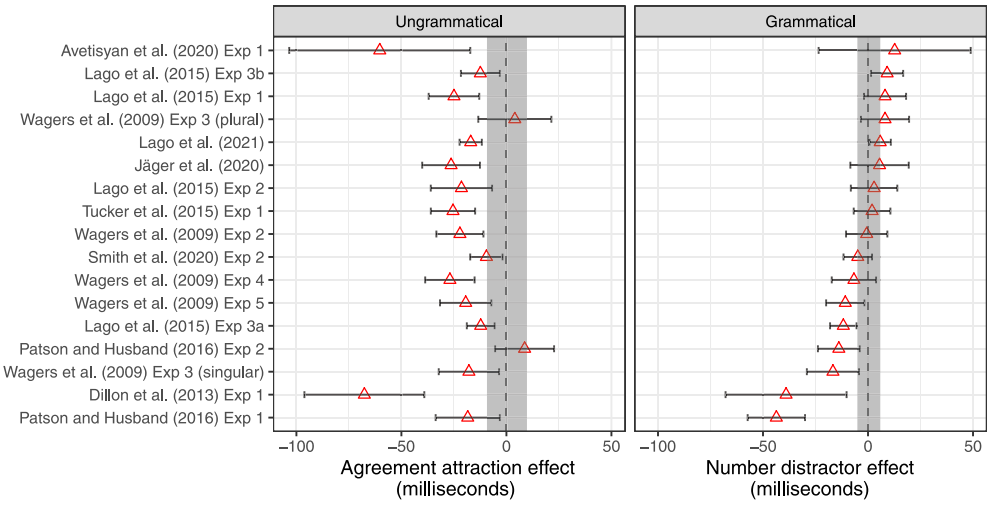


Fig. 4. Prior predictions of the non-linear cue-based retrieval model: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model. The red triangles and the error bars around them show the observed effects for each dataset specified on the *y*-axis.

Table 1
The percentage of agreement errors observed in acceptability-judgment studies. The last column named 'Agreement error due to number attraction' reports the difference in percentage of agreement errors between the agreement attraction condition e.g., *the key to the cabinets were..* and the baseline condition e.g., *the key to the cabinet were...* The values in square brackets represent 95% confidence intervals.

Experiment	Number of subjects	Language	Agreement error (in percentage) due to agreement attraction
(Häussler, 2009) Exp 1	48	German	2 [−2, 6]
(Häussler, 2009) Exp 2	32	German	17 [8, 26]
(Häussler, 2009) Exp 3	64	German	15 [9, 21]
(Häussler, 2009) Exp 5	40	German	20 [11, 29]
(Häussler, 2009) Exp 6 (Adjacent RC)	40	German	12 [4, 20]
(Häussler, 2009) Exp 6 (Non-Adjacent RC)	40	German	8 [2, 14]
(Hammerly et al., 2019) Exp 1	40	English	20 [10, 30]
(Hammerly et al., 2019) Exp 2	20	English	24 [7, 41]
(Hammerly et al., 2019) Exp 3	40	English	16 [7, 25]
(Laurinavichyute & von der Malsburg, 2022) Exp 1	1072	English	21 [12, 30]
(Schlueter et al., 2018) Exp 1	30	English	28 [15, 41]
(Schlueter et al., 2018) Exp 3	30	English	38 [26, 50]
(Schlueter et al., 2018) Exp 4	30	English	24 [10, 38]
(Tanner et al., 2014) Exp 1	24	English	11 [−2, 24]
(Tanner et al., 2014) Exp 2	22	English	12 [3, 21]

sentence (a) is always rated more acceptable than sentence (b). The higher acceptability ratings in (a) are often taken as evidence of a change in the representation of the subject noun (Bock & Eberhard, 1993; Eberhard, Cutting, & Bock, 2005; Hammerly et al., 2019): for example, the plural feature on the non-subject noun may overwrite the feature representation of subject noun phrase in a proportion of trials. Under the assumption that a change in the representation of subject noun drives the higher acceptability in sentences like (a), one can infer that the difference in acceptability between sentence (a) and (b) reflects the rate of distortion in the system. The difference in agreement errors has been found to mostly occur in the range of 10% to 30% across five acceptability judgment studies as shown in Table 1 (with a meta-analytical estimate of 12%–22%).[7] Given this observed range of apparent agreement attraction errors, a prior distribution that constrains the distortion rate to lie between approximately 10% and 50% would be a reasonable one. We use such a prior on the distortion rate parameter across all the models that assume some kind of representation distortion.

[7] For the details of the meta-analysis, see section 5.2.1 of the supplementary workflow document here: https://osf.io/gqj3p/

The feature percolation model

The feature percolation model (Bock & Eberhard, 1993; Eberhard, 1997)[8] was proposed exclusively for subject–verb number agreement dependencies. The model assumes that

(1) The plural feature of the distractor noun percolates to the subject noun in a proportion of trials, causing a change in the feature representation of the subject.
(2) Dependency completion is faster when the subject noun matches in number feature with the verb compared to when it does not.

Assumption (2) is not a part of the original proposals about feature percolation. We make this additional assumption in order to generate reading times from the model. The rationale for this assumption is that when the subject noun matches in number feature with the verb, it can license the number marking on the verb, resulting in easier

[8] The feature percolation model is in the broader category of encoding-based models. Another well-known encoding-based model is the Marking and Morphing model (Eberhard et al., 2005). Both the models make almost similar predictions about agreement attraction in reading; see section The Marking and Morphing model.

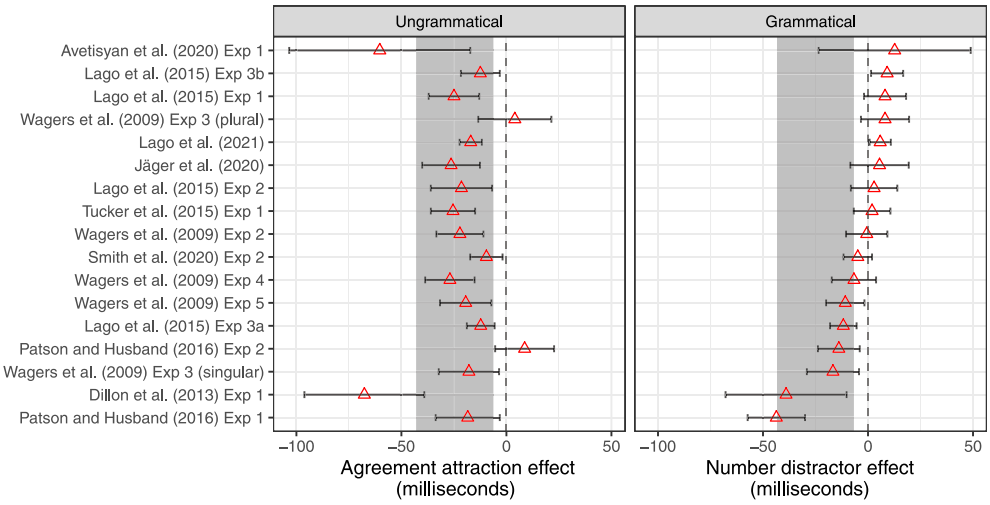


Fig. 5. Prior predictions of the feature percolation model: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model. The red triangles and the error bars around them show the observed effects for each dataset specified on the *y*-axis.

processing compared to when the verb's number is not licensed by the subject (Pearlmutter, Garnsey, & Bock, 1999).

Consider again the ungrammatical conditions (c) and (d) in Fig. 2. The feature percolation model proposes that the [+plural] feature of *the cabinets* in condition (c) percolates up to the subject *the key* in a proportion of trials and changes the number of the subject noun from singular to plural. Consequently, the plural subject is able to license the plural verb in those trials, causing faster reading times on average compared to condition (d) where no feature percolation happens. Thus, in ungrammatical sentences, the model predicts the agreement attraction effect—faster reading times in condition (c) compared to (d). In the grammatical conditions (a) and (b) in Fig. 2, the model predicts faster reading times in (a) vs. (b) because the plural feature on the distractor in condition (b) percolates up to the subject noun, causing a disruption in the licensing of the singular verb in a proportion of trials. See section The feature percolation model for the implementational details of the model.

We generate prior predictions from the model assuming two free parameters, a scaling parameter S and a distortion rate parameter θ. The scaling parameter determines the range of reading times generated by the model. The prior on the scaling parameter S is a truncated normal distribution with mean 5.3, standard deviation 0.05, and a lower bound at 5.2. Similar to the cue-based retrieval model, we choose this lower bound on scaling because the model generates reasonable reading times under this constraint (see section Choice of priors on the scaling parameter).

$$S \sim Normal_{lb=5.2}(5.3, 0.05) \tag{3}$$

The distortion rate parameter is the probability with which the plural feature percolates to the subject noun. The higher the feature distortion rate, the larger the effects in both grammatical and ungrammatical conditions. Because the distortion rate modulates the size of the attraction effect predicted by the model, and because the data show considerable variability, we treat this parameter as a free parameter and estimate it from the observed effects. We choose a truncated-normal prior on the distortion rate θ, so that the degree of distortion is constrained between 10% and 50%:

$$\theta \sim Normal_{lb=0.1}(0, 0.25) \tag{4}$$

The term $lb = 0.1$ indicates a lower bound on distortion rate values. The lower bound of 0.1 implies that the value of distortion rate lies in the range [0.1, 0.5] — that is, the representation of the subject noun changes in at least 10% and at most 50% of the trials.[9] The prior predictions of the model given the above priors are shown in Fig. 5. The model correctly captures the observed agreement attraction effects (in ungrammatical sentences) but it fails to entirely capture the number distractor effect (in grammatical sentences).

The grammaticality bias model

In a recent paper, Hammerly et al. (2019) claimed that the asymmetry of attraction effects in grammatical and ungrammatical sentences is due to a phenomenon called response bias. Their claim relates to grammaticality-judgment data, where participants have to judge whether a given sentence is grammatical or ungrammatical. A consistent finding from judgment studies is that of agreement attraction: the participants make more incorrect and slower judgments in sentences where a non-subject noun matches in number with the verb compared to sentences where it does not. However, an asymmetry is observed in judgment data between grammatical and ungrammatical sentences: the attraction effects are relatively small in grammatical sentences compared to ungrammatical sentences. Hammerly and colleagues explain this asymmetry in acceptability judgments using a response bias parameter such that the participants are biased to respond 'grammatical' in judgment tasks. This response bias assumption (implemented in a drift–diffusion model) explains the smaller effect sizes in grammatical sentences.

The authors speculate that the response bias proposal can be extended to reading time studies and it can potentially explain the asymmetry of number agreement effects in reading. To extend their proposal for reading times, we implement a marking and morphing model (Eberhard et al., 2005) assuming a *grammaticality bias*: the comprehender has a strong expectation that the partially-seen sentence will be followed by a grammatical continuation. For example, in the case of subject–verb agreement dependencies, there is a bias that the noun phrases will be followed by a verb that agrees in number with the subject noun. Consequently, any increase in number-mismatch between the subject and verb causes an exponential increase in the processing difficulty at the verb. We operationalize this grammaticality bias idea in the marking and morphing model as follows.

[9] The upper bound of 0.5 is not a hard constraint; it is an approximate upper bound that arises due to the normal prior with mean 0 and standard deviation 0.25 (Johnson, Kotz, & Balakrishnan, 1995).

The marking and morphing model assumes that the subject noun has a continuous-valued representation of number such that an unequivocally singular noun will have the lowest value and an unequivocally plural noun will have the highest value. In addition, the plurality of the non-subject nouns can spread to the subject noun to make it more plural. This continuous-valued plurality of the subject noun determines the processing difficulty at the verb such that a number-mismatch between the subject and the verb would produce a processing cost at the verb (on a continuous scale), which we call the *mismatch cost*. For example, if $S(r)$ is the value of plurality of the subject and V_{pl} is the value of plurality of the verb, the mismatch cost at the verb Δ is given by

$$\Delta = \left| S(r) - V_{pl} \right| \cdot \delta \tag{5}$$

where δ is a constant used for scaling the degree of mismatch $|S(r) - V_{pl}|$ on the log milliseconds scale; $S(r)$ represents the continuous-valued number of the subject noun (scaled between 0 and 1 such that 0 indicates unequivocally singular and 1 indicates unequivocally plural); V_{pl} indicates the plurality of the verb such that the value of V_{pl} is 1 if the verb is plural and 0 if the verb is singular.

In order to implement the grammaticality bias idea, we introduce a bias parameter in the above equation such that an increase in degree of mismatch between the subject and the verb should cause an exponential increase in the mismatch cost. The modified equation with the grammaticality bias parameter b is given by

$$\Delta = \left(\left| S(r) - V_{pl} \right| \cdot \delta \right)^{2b} \tag{6}$$

where the bias parameter b can take values between 0.5 and 1. A value of 0.5 would mean there is no grammaticality bias and the model would be the same as the default marking and morphing model, while $b = 1$ indicates the strongest grammaticality bias.

The above equation implies that the grammaticality bias parameter minimizes the impact of a small mismatch between the subject and the verb's number (as in grammatical sentences) but maximizes the impact of a large mismatch (as in ungrammatical sentences) on the processing cost at the verb. As the value of this parameter increases, the asymmetry between grammatical and ungrammatical sentence increases. As implemented in this paper, what the grammaticality bias model achieves is to formalize the extent to which a modest feature mismatch is tolerated as a grammatical continuation, and the extent to which a large feature mismatch is penalized during reading. Thus, unlike the response bias parameter in Hammerly et al.'s model, our grammaticality bias parameter does not directly represent a bias in reading grammatical vs. ungrammatical sentences. Rather it reflects a bias in the internal cognitive process that computes the cost of feature mismatch between two co-dependents during reading: a minor feature mismatch has a negligible effect on processing but a large mismatch has a penalizing effect on processing. However, our grammaticality bias proposal is similar to the response bias proposal of Hammerly et al. in that it predicts an asymmetry in number agreement effects between grammatical and ungrammatical sentences.

Fig. 6 illustrates the effect of grammaticality bias on processing cost at the verb in grammatical and ungrammatical sentences. The difference in mismatch cost between sentences (a) and (b) determines the magnitude of the number distractor effect in grammatical sentences and the difference between (c) and (d) determines the magnitude of the agreement attraction effect in ungrammatical sentences. The figure shows that as the grammaticality bias increases (e.g., $b = 1$), the asymmetry between the grammatical and the ungrammatical sentences increases while the effect is symmetrical when there is no bias i.e. when b is equal to 0.5.

To generate reading time predictions at the verb, we use a lognormal distribution. The reading time in the kth trial as a function of the number valuation of the subject noun and the grammaticality bias parameter is given by

$$T_k \sim Lognormal\left(\alpha + \left(\left| S(r) - V_{pl,k} \right| \cdot \delta \right)^{2b}, \sigma \right) \tag{7}$$

where α is the scaling parameter, it can be interpreted as the mean reading time (in log ms) when the noun number exactly matches the verb number; $S(r)$ represents the continuous-valued number of the subject noun (scaled between 0 and 1), $V_{pl,k}$ indicates plurality of the verb in trial k (0 for singular, 1 for plural), and b indicates the grammaticality bias parameter. See section The grammaticality bias model for the implementational details of the model.

Fig. 7 demonstrates the prior predictions of the model with and without the grammaticality bias. The model without any grammaticality bias becomes the default marking and morphing model and predicts the effect sizes similar to the feature percolation model. But when the model has a grammaticality bias, of say 0.75, it predicts close-to-zero effects in grammatical sentences. The model is thus better than the feature percolation model at capturing the number distractor effect in grammatical sentences but it still fails to capture the entire range, especially the positive effects, in grammatical sentences.

The lossy compression model

Another type of representation distortion-based model is based on the *lossy compression of the linguistic input*. Here, the assumption is that the feature representation of the linguistic input changes probabilistically when it is stored in memory. The comprehender thus has access to only a potentially distorted memory representation of the true intended message. The representation distortion in these models is constrained by information-theoretic principles. One model of this type is the lossy-context surprisal model of Futrell et al. (2020) which has been shown to explain structural forgetting effects in English and German (Gibson & Thomas, 1999; Vasishth, Suckow, Lewis, & Kern, 2010) and information locality across languages (Futrell, 2019).

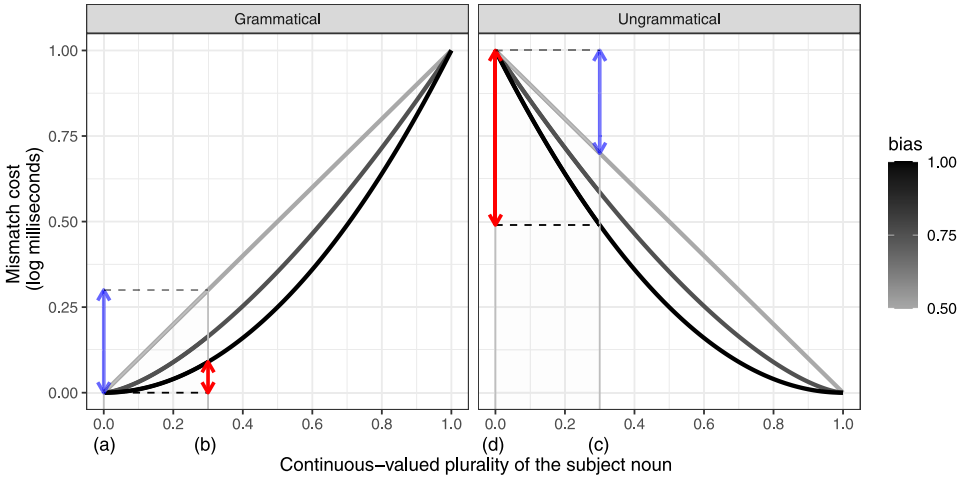
We implement a lossy compression model of number agreement effects in subject–verb number agreement dependencies. The model makes the following assumptions:

(1) The actual representation of the pre-verbal input gets distorted to an imperfect memory representation for the comprehender such that the plural marker on the nouns can be inserted or deleted at constant rates.
(2) From the possibly-distorted memory representation, the comprehender reconstructs a set of possible true representations conditional on their prior linguistic knowledge and their uncertainty about the degree of distortion.
(3) The processing difficulty at the verb is the expected surprisal – the negative log probability – of encountering the verb given all possible memory representations of the pre-verbal input.[10]

Consider the grammatical sentence *The key to the cabinets was rusty*. Here the linguistic material preceding the verb is a noun phrase modified by a prepositional phrase and contains two nouns, *key* and *cabinets*. The first noun is singular and the second noun is plural. The observed input thus has the representation **NPN**.**pl**, where the first **N** represents the singular head noun of the phrase, **P** represents the preposition, and **N**.**pl** represents the plural-marked noun inside the prepositional phrase.

The lossy compression model assumes that the comprehender has access only to a memory representation of the observed input **NPN**.**pl**. For instance, it is possible that the input distorts to **N**.**plPN**.**pl**, where a plural marker is inserted at the first noun due to lossy memory. There are four such possible memory representations that can arise

[10] The probability of seeing the verb given a memory representation is calculated by marginalizing out the possible true representations; see section The lossy compression model .



(a) The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{+singular}$ $was^{subject}_{singular}$ rusty

(b) The $key^{+subject}_{+singular}$ to the $cabinets^{-subject}_{-singular}$ $was^{subject}_{singular}$ rusty

(c) * The $key^{+subject}_{-plural}$ to the $cabinets^{-subject}_{+plural}$ $were^{subject}_{plural}$ rusty

(d) * The $key^{+subject}_{-plural}$ to the $cabinet^{-subject}_{-plural}$ $were^{subject}_{plural}$ rusty

Fig. 6. An illustration of *mismatch cost* at the verb as a function of subject's plurality and the grammaticality bias. The labels (a), (b), (c), and (d) on the x-axis mark the number value of the subject noun in four conditions shown below the graph. The red arrows – indicating the difference in mismatch costs between a pair of conditions – can be interpreted as the number agreement effects on the log scale when there is a grammaticality bias ($b = 1$); similarly, the blue arrows indicate agreement effects (on log scale) when there is no bias ($b = 0.5$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

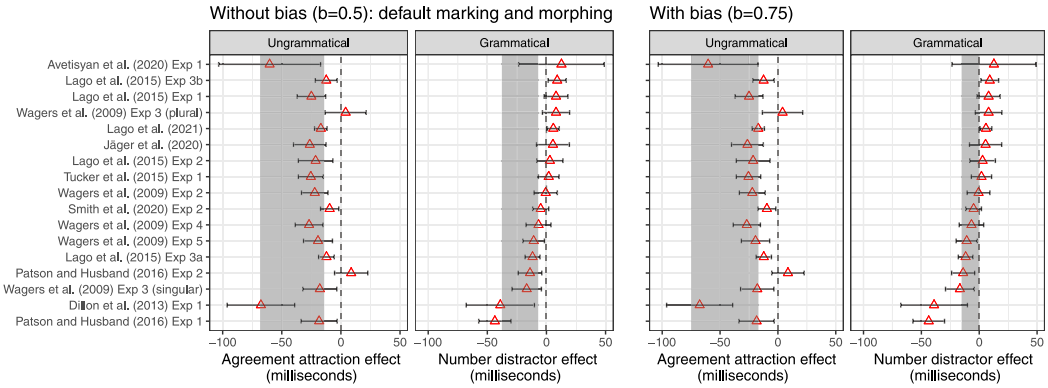


Fig. 7. Prior predictions of the grammaticality bias model: The shaded gray bands represent the 95% credible interval of attraction effect predicted by the model. The red triangles and the error bars around them show the observed effects for each dataset.

The key to the cabinets **was** rusty

I = N P N.pl

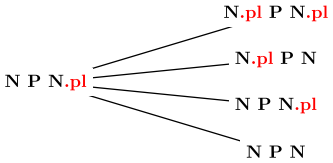


Fig. 8. An schematic illustration of lossy compression mechanism: a given input **I = NPN.pl** can distort to the four possible memory representations due to insertion and/or deletion of plural markers on the nouns.

due to insertion or deletion of plural morphemes when constructing the memory representation (see Fig. 8).

For each possible memory representation, one can calculate the conditional probability of seeing a verb with a particular number marking given that memory representation. Lossy compression assumes that the processing difficulty at the verb is proportional to the average of the negative logarithm of these conditional probabilities. In other words, how hard a verb is to process is related to how unexpected the verb would be given all of the different ways the preceding noun phrase might be misremembered. A detailed discussion of the model's implementation can be found in section The lossy compression model .

We generate the prior predictions from the model assuming three free parameters, the scaling parameter S, and two distortion rate parameters a and d, where a represents the rate of inserting a plural marker and d represents the rate of deleting a plural marker. The scaling parameter determines the slope of the linear function that links the processing difficulty to the reading times at the verb. For the prior on the scaling parameter S, we used a truncated normal distribution,

$$S \sim Normal_{lb=0.15}(0.25, 0.05) \tag{8}$$

where $lb = 0.15$ indicates a lower bound of 0.15 on scaling values.

The two distortion rate parameters, the insertion rate a and the deletion rate d, determine the degree of information loss when the intended message transforms to a memory representation. We set the same prior on distortion rate parameters a and d as we did for distortion rate in the feature percolation model.

$$a, d \sim Normal_{lb=0.1}(0, 0.25) \tag{9}$$

where $lb = 0.1$ indicates a lower bound of 0.1 on insertion and deletion rate values.

Two other parameters in the model – the prior knowledge about the possible pre-verbal input and the surprisal at the verb – are estimated from the corpus data (Nivre, Abrams, et al., 2018; Schäfer, 2015; Schäfer & Bildhauer, 2012) and their values differ across different experimental designs and languages. Therefore, the model generates slightly different predictions for the eight different experimental designs (which include different languages). The prior predictions of the model given the above priors are shown in Fig. 9. The observed number distractor effects (in grammatical sentences) are mostly consistent with the model's predictions but the number agreement effects are not fully captured by the model.

Two hybrid representation distortion-plus-retrieval models

We propose a new class of models that assumes a hybrid mechanism combining representation distortion- and retrieval-based processes: the representation of the pre-verbal linguistic material can get distorted before the retrieval is triggered at the verb. The content-addressable search that happens during retrieval now involves potentially distorted noun representations instead of a veridical representation of the input. We propose two models of this type—the first model combines feature percolation and cue-based retrieval, and the second model combines lossy compression and cue-based retrieval. We discuss these next.

The feature percolation-plus-retrieval model

The feature percolation-plus-retrieval model proposes that the representation of the subject noun changes due to feature percolation before retrieval is triggered at the verb. The model assumes that

(1) Dependency completion between the subject and the verb is driven by cue-based retrieval.
(2) The retrieval at the verb is preceded by a probabilistic feature percolation from the distractor noun to the subject noun.

Consider the grammatical conditions (a) and (b) shown in Fig. 10. The distractor noun in condition (b) has a plural feature that can percolate up to the subject noun and change its representation. Suppose that the rate of feature percolation is θ. The plural feature would percolate up to the subject noun in θ proportion of trials. If there were a total of N trials, the subject noun in condition (b) will have a plural feature in $\theta \times N$ number of trials and a singular feature in $(1-\theta) \times N$ trials. In the $\theta \times N$ trials in which the subject noun becomes plural, processing will be slowed during retrieval because the subject is no longer a full feature match for the verb's retrieval cues. In the $(1-\theta) \times N$ trials where no percolation occurs, retrieving the unmodified subject happens quickly because it is a good feature match with the verb.

By contrast, in condition (a), there is no plural feature in the input that could percolate up to the subject, so the subject's feature representation remains intact in all N trials. This makes the subject a good feature match with the verb, but processing is still slowed due to the fan effect: The singular distractor is a partial feature match with the verb and thus receives some activation that would have otherwise gone to the subject, slowing processing. As a result, depending on the value of θ, the model can predict facilitation, inhibition, or no effect in the grammatical conditions when considering the difference between (a) and (b). In the ungrammatical conditions, the model always predicts a facilitatory effect regardless of the value of θ. See section The feature percolation-plus-retrieval model for a detailed explanation of the model's implementation and predictions.

We generate prior predictions from the model assuming two free parameters, the scaling parameter F, and the distortion rate parameter θ. The scaling parameter serves the same function as in the cue-based retrieval model. So we choose the same prior on the scaling parameter as in the cue-based retrieval model,

$$F \sim Normal_{lb=0.05}(0.15, 0.05) \tag{10}$$

For the distortion rate parameter θ – which represents the probability of feature percolation – we choose the same prior as in the feature percolation model,

$$\theta \sim Normal_{lb=0.1}(0, 0.25) \tag{11}$$

Here, as discussed earlier, $lb = 0.1$ indicates a lower bound of 0.1 on distortion rate values. The prior predictions of the model given the above priors are shown in Fig. 11. The model predictions are largely consistent with observed effects in both grammatical and ungrammatical sentences. To our knowledge, this is the first implemented computational model that captures the qualitative pattern observed in published subject–verb number agreement studies on sentence comprehension.

The lossy compression-plus-retrieval model

Another model we propose is the lossy compression-plus-retrieval model. Here, the idea is that the representation of the pre-verbal input changes due to lossy compression before the retrieval is triggered at the verb. The model assumes that

(1) Dependency completion between the subject and the verb is driven by cue-based retrieval.
(2) The retrieval at the verb is preceded by the distortion of the pre-verbal input to an imperfect memory representation such that the plural marker on the nouns can be deleted or inserted.

The above assumptions imply that the retrieval process takes place on a potentially-distorted memory representation of the nouns, which affects the activation received by each noun and consequently, the retrieval times at the verb. For example, in the sentence *The key to the cabinets was rusty*, if the representation of preverbal nouns remains intact, then the subject noun *the key* receives activation from both [subject] and [singular] cue at the verb. But if the input distorts to an imperfect memory representation, say *the keys to the cabinets...*, the subject noun then receives activation only from the [subject] cue at

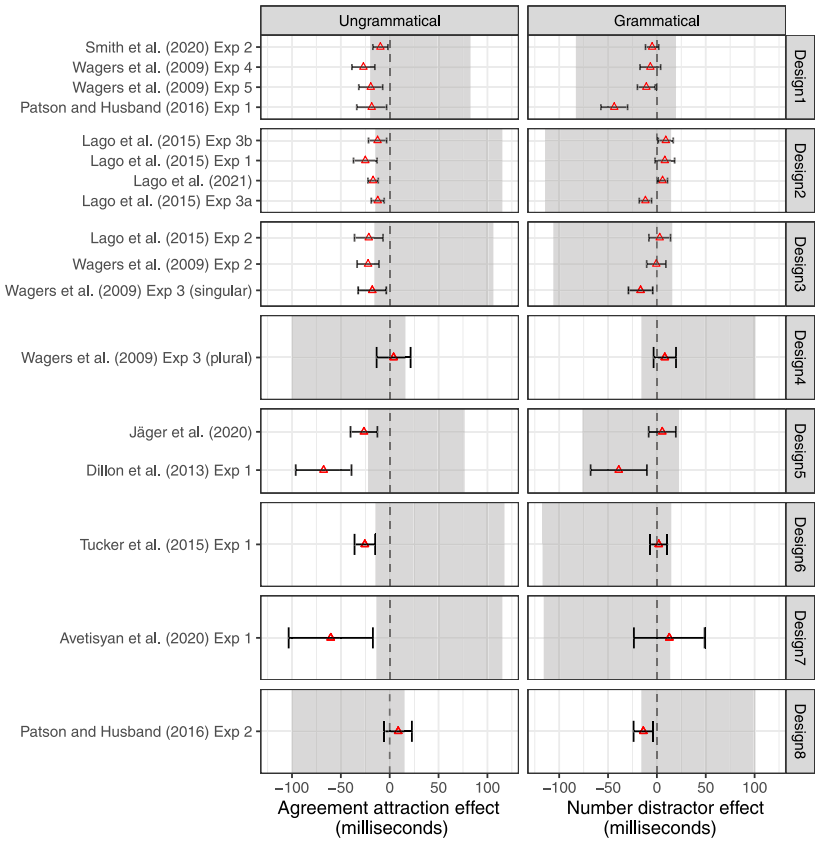


Fig. 9. Prior predictions of the lossy compression model: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model; the predicted range of effect differs across experimental designs involving subject–verb number agreement. Because the lossy compression model's predictions depend on corpus data, the different designs and languages used in the experiments need to be differentiated. Design 1 refers to English prepositional phrase constructions with singular subject noun; Designs 2 and 7 refer to relative clause constructions in Spanish and Armenian respectively; Designs 3 and 4 refer to English object relative clause constructions with singular and plural subject respectively; Designs 5 and 6 refer to subject relative clause constructions in English and Arabic respectively; Design 8 refers to English prepositional phrase constructions where the subject noun's number is manipulated. The red triangles and the error bars around them show the observed number agreement effects for each dataset specified on the y-axis.

Table 2
The priors on the scaling parameter and the feature distortion rate parameter for each model used in the evaluation; *Normal* represents a normal distribution, *lb* stands for lower bound.

Model	Prior on scaling parameter	Prior on distortion rate
Cue-based retrieval model	$Normal_{lb=0.05}(0.15, 0.05)$	–
Non-linear cue-based retrieval model	$Normal_{lb=5.2}(5.3, 0.05)$	–
Feature percolation model	$Normal_{lb=5.2}(5.3, 0.05)$	$Normal_{lb=0.1}(0, 0.25)$
Grammaticality bias model	$Normal_{lb=5.2}(5.3, 0.05)$	$Normal_{lb=0.1}(0, 0.25)$
Lossy compression model	$Normal_{lb=0.15}(0.25, 0.05)$	$Normal_{lb=0.1}(0, 0.25)$
Feature percolation-plus-retrieval model	$Normal_{lb=0.05}(0.15, 0.05)$	$Normal_{lb=0.1}(0, 0.25)$
Lossy compression-plus-retrieval model	$Normal_{lb=0.05}(0.15, 0.05)$	$Normal_{lb=0.1}(0, 0.25)$

the verb. The implementational details of the model are discussed in section The lossy compression-plus-retrieval model.

We generate prior predictions from the model assuming three free parameters, the scaling parameter F and two distortion rate parameters, insertion rate a, and deletion rate d. We choose the same priors on the scaling and distortion rate as in the feature percolation-plus-retrieval model. The prior predictions of the model given these priors are shown in Fig. 12. As with the lossy-compression model, the prior predictions depend on the experimental design in each study.

Table 2 shows the priors on the scaling and the distortion rate parameter for each of the above-discussed five models. Next, we compare the predictive accuracies of these five models on the data from published studies on subject–verb number agreement.

Model comparison

To quantify the evidence for each model, we estimate their predictive accuracies on the observed data using cross-validation. The data consist of agreement attraction and number distractor effect estimates from the 17 published datasets on subject–verb number agreement studies (Avetisyan et al., 2020; Dillon et al., 2013; Jäger et al., 2020; Lago, et al., 2015; Patson & Husband, 2016; Smith, Franck, & Tabor, 2021; Tucker et al., 2015; Wagers et al., 2009).[11] The effects in

[11] To estimate number agreement effects for each dataset, we fit a Bayesian linear-mixed model with the main effect of grammaticality and the nested

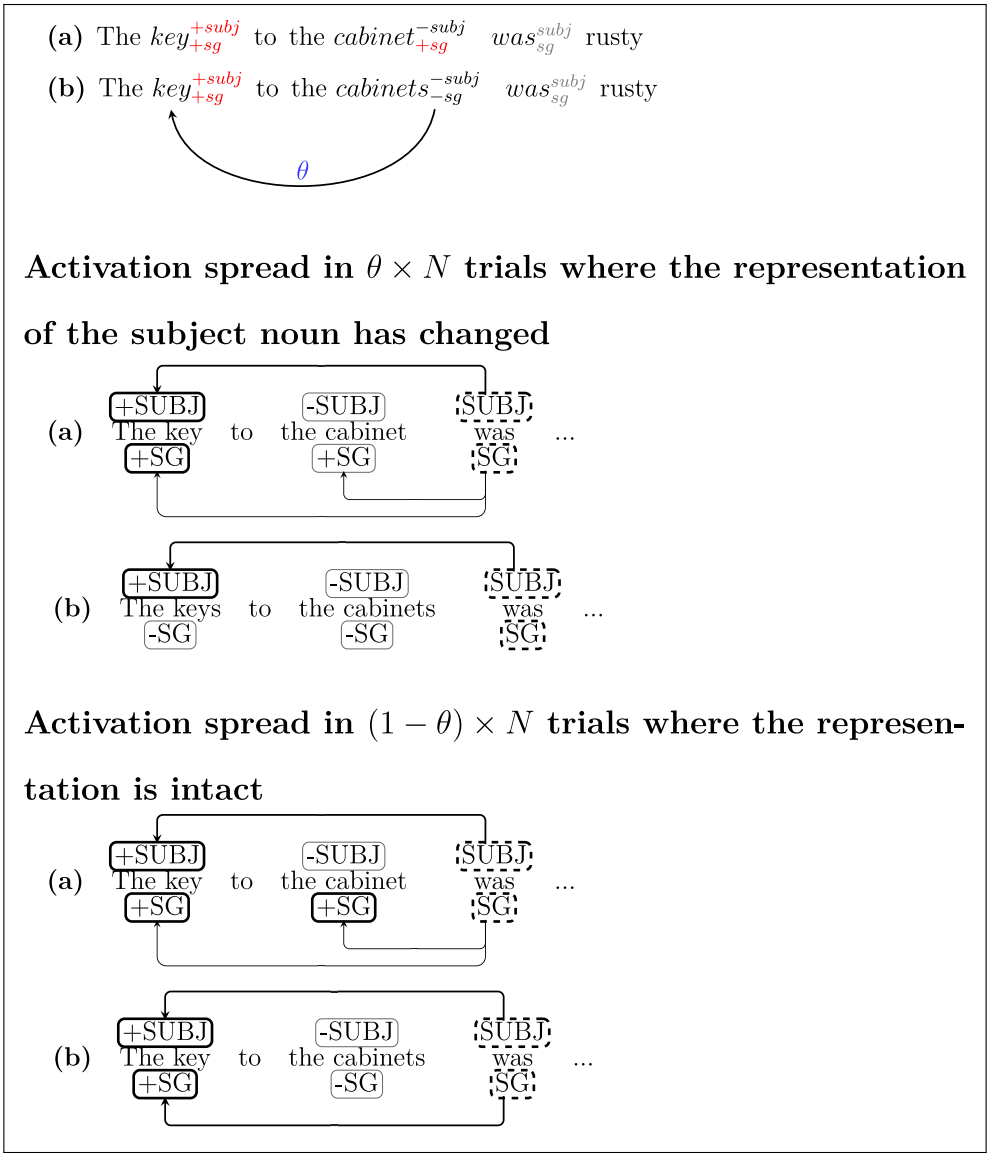


Fig. 10. Schematic illustration of activation received by the nouns in the feature percolation-plus-retrieval model: Feature percolation modulates the amount of activation received by the subject noun during the retrieval process. In $\theta \times N$ trials, the subject noun receives less activation in condition (b) compared to condition (a).

grammatical and ungrammatical conditions from these 17 datasets are shown in Fig. 1.

We use cross-validation for estimating predictive accuracies. Cross-validation allows us to compute how accurately a model performs on a portion of the data after being trained on a different portion of the same data. We implement cross-validation as follows. First, 17 sets of training and test data are created by leaving out one dataset as the test data and taking the other 16 as training data. Second, in each iteration i, the models are fitted on training data $D_{\text{train},i}$ and their predictive accuracies are computed on the test data $D_{\text{test},i}$.[12] The code and data are available from https://osf.io/gqj3p/.

For model fitting, we use approximate Bayesian computation (ABC) (Palestro, Sederberg, Osth, Van Zandt, & Turner, 2018; Sisson, Fan, & Beaumont, 2018). ABC compares the training data and the model-simulated data to infer what values of the parameter(s) would have generated the given training data. A simple ABC algorithm works as follows: (i) a parameter value say Θ^* is sampled from the prior distribution, (ii) data is generated from the model conditional on the sampled value Θ^*, i.e., by plugging Θ^* into the model, (iii) if the model-generated data is *close enough* to the training data, Θ^* is accepted as a sample from the posterior distribution. See section Parameter

effects of attractor type (matching vs. mis-matching attractor) within grammatical and ungrammatical conditions; the dependent variable was reading time in milliseconds.

[12] Each training set contained 16 × 2 data points — the mean estimates of agreement attraction and number distractor effects from 16 (training) datasets.

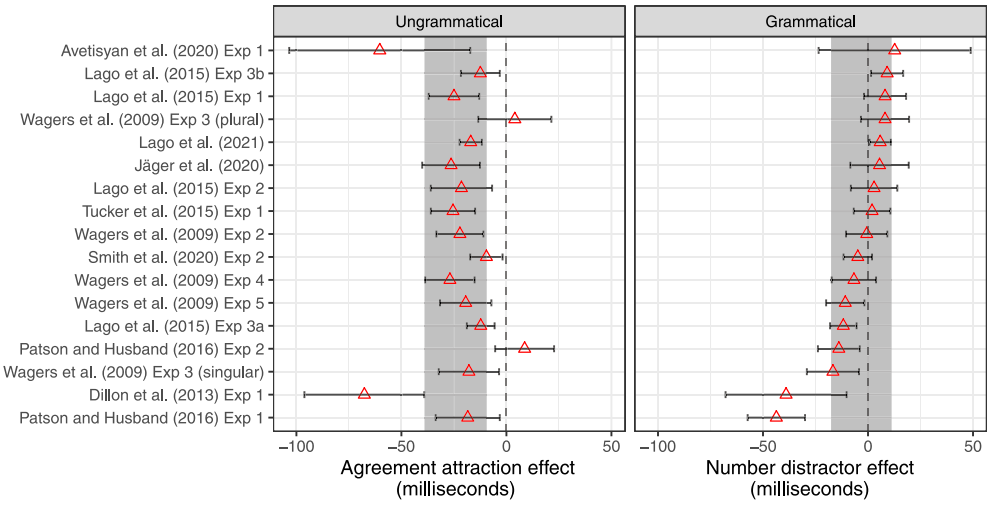


Fig. 11. Prior predictions of the feature percolation-plus-retrieval model: The shaded gray bands represent the 95% credible interval of attraction effect predicted by the model. The red triangles and the error bars around them show the observed effects for each dataset.

estimation for details of the implementation of ABC that we used. ABC allows us to fit a model using the Bayesian method without requiring us to define an explicit likelihood function. We use this method because parameter estimation becomes computationally infeasible if we use the exact likelihood function for three of the models we evaluated here, the cue-based retrieval model and the two hybrid models. A further motivation for using ABC is that the framework we have developed in this paper can be used for any process model of sentence processing that makes reading time predictions, even if no likelihood function can be derived for the model. Examples of such models are the dynamical systems models discussed in Engbert, et al. (2022), Rabe, et al. (2021), Smith et al. (2021), and Smith and Vasishth (2022).

The model fit using ABC provides the posterior distributions of the parameters of a model. To compute the predictive accuracy of a model, parameter values are sampled repeatedly from the posterior distribution of each parameter. Then, the likelihood of generating the test data given each set of sampled parameter values is computed. The predictive accuracy of a model on the test data $D_{\text{test},i}$ is the log of likelihoods averaged over the posterior distribution estimated from the training data $D_{\text{train},i}$. A model's overall predictive accuracy, the expected log predictive density ($\widehat{elpd}$), is computed as the sum of its log predictive densities over 17 iterations along with the standard error of sum. See section Cross-validation method for a detailed note on the method. We get a measure of predictive performance for each model — the $\widehat{elpd}$ value. From the $\widehat{elpd}$ values, we can compute the difference in performance of each pair of models, represented by $\Delta\widehat{elpd}$.

Results

Fig. 13 shows the $\Delta\widehat{elpd}$ values – the difference in predictive performance – for each pair of models along with the standard error (SE) of difference. The $\Delta\widehat{elpd}$ value represents the strength of evidence in favor of one model over another. The $\Delta\widehat{elpd}$ values are interpreted as follows. If the $\Delta\widehat{elpd}$ for a pair of models is larger than $2\times SE$, then the two models are distinguishable in their performance and one of them has clear evidence in its favor over the other.

In Fig. 13, the positive difference in $\widehat{elpd}$ values implies that the model shown in a graph's title performs better than the other model. A pair of models is distinguishable if the error bar does not cross the zero line (where $\Delta\widehat{elpd} = 0$). The $\Delta\widehat{elpd}$ analysis reveals four key results.

1. The feature percolation-plus-retrieval model shows positive $\Delta\widehat{elpd}$ values when compared with each of other six models. But the hybrid model is indistinguishable from the feature percolation model and the grammaticality bias model. This implies that the feature percolation-plus-retrieval model outperforms all other models except the feature percolation and the grammaticality bias model, with which the hybrid model has comparable performance.
2. The feature percolation model and the grammaticality bias model show mostly positive $\Delta\widehat{elpd}$ values against the lossy compression model and the lossy compression-plus-retrieval model. But the error bars around the $\Delta\widehat{elpd}$ values always cross zero, implying that there is no clear evidence in the favor of the feature percolation and the grammaticality bias model over the two lossy compression-based models.
3. The lossy-compression-plus-retrieval model shows positive $\Delta\widehat{elpd}$ values against the lossy compression model and two cue-based retrieval models, indicating decisive evidence in favor of this hybrid model over the lossy compression and cue-based retrieval models.
4. The non-linear cue-based retrieval model performs better than the cue-based retrieval model of Lewis and Vasishth (2005) (positive $\Delta\widehat{elpd}$ value) and performs almost similar to the lossy-compression model ($\Delta\widehat{elpd}$ value is close to zero with large standard errors); the model shows negative $\Delta\widehat{elpd}$ values against the all other distortion-based and hybrid models indicating that it is one of the worst performing models.
5. The cue-based retrieval model shows negative $\Delta\widehat{elpd}$ values against the other six models, meaning that the cue-based retrieval model has the worst predictive performance of all the models considered. However, the error bars around the $\Delta\widehat{elpd}$ values indicate that the model's performance is not distinguishable from the feature percolation and the lossy compression model.

A model's predictive performance can in principle be sensitive to the choice of priors on the parameter of interest (Schad, Nicenboim, Bürkner, Betancourt, & Vasishth, 2022); here, this parameter is the distortion rate. To understand the impact of prior specification, we also evaluated each model's performance under different prior assumptions about the degree of distortion in the system. The overall pattern of results remains the same; only the feature percolation model and the grammaticality bias model's performance slightly decreases with an

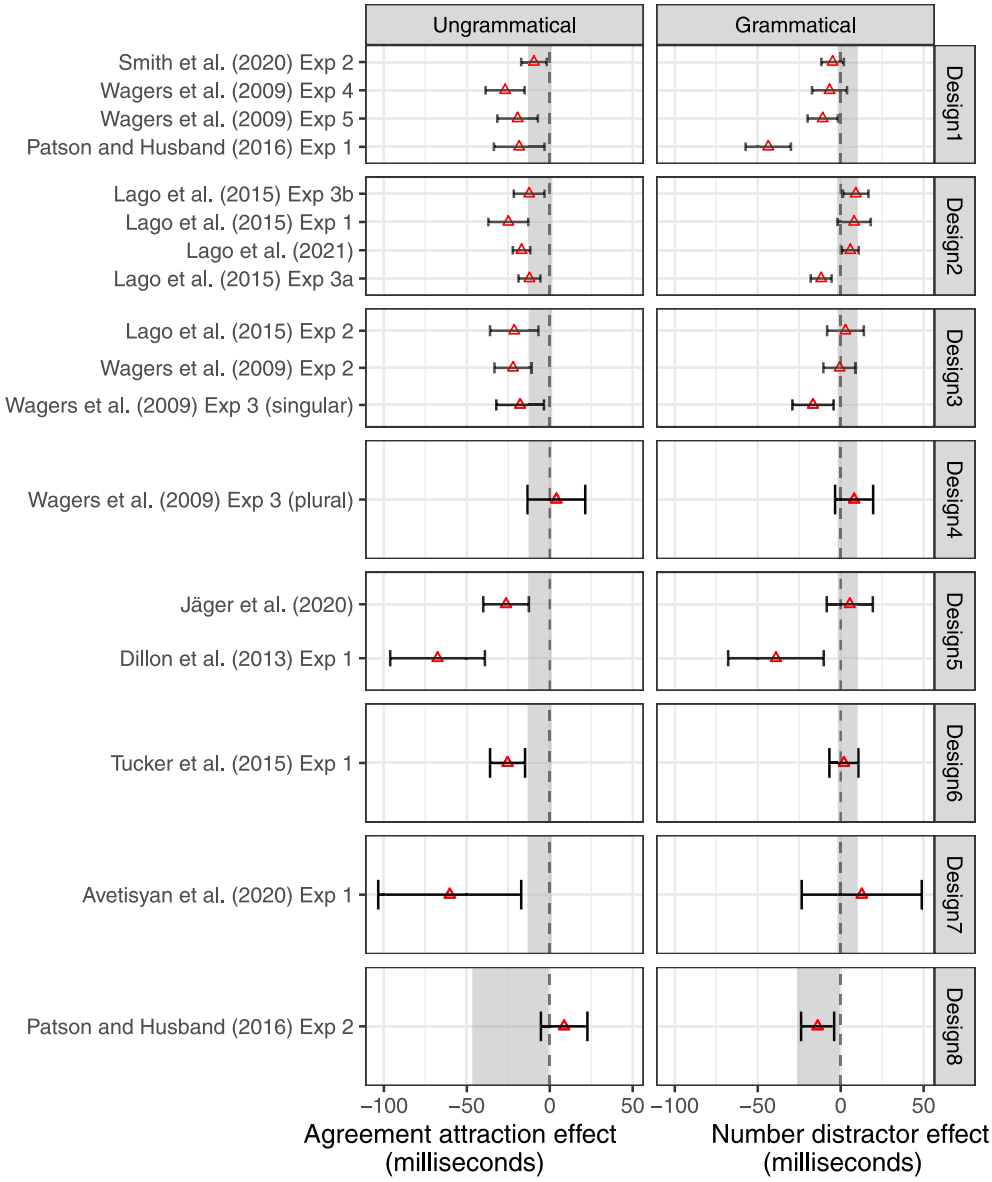


Fig. 12. Prior predictions of the lossy compression-plus-retrieval model: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model; the predicted range of effect differs across experimental designs. The red triangles and the error bars around them show the observed effects for each dataset specified on the y-axis.

increase in lower bound on distortion rate. For example, under the assumption that the distortion rate is higher than 0.25, both the feature percolation model and the grammaticality bias model perform (distinguishably) worse than the hybrid percolation-plus-retrieval model. The prior sensitivity analysis is shown in section Prior sensitivity analysis.

Discussion

The model comparison shows that, among the seven models considered here, the hybrid feature-percolation-plus-retrieval model shows the best performance numerically in explaining the observed patterns across the subject–verb number agreement datasets. As Fig. 13 shows, the hybrid model decisively outperforms all the other models except the feature-percolation model and the grammaticality bias model.[13] The feature percolation and the grammaticality bias models do slightly better than the two lossy compression models and the cue-based retrieval models, but this improvement in fit in these models is not convincing. The lossy compression-plus-retrieval model outperforms the lossy compression and the cue-based retrieval models. Finally, the cue-based retrieval model has the worst performance of all the models considered.

[13] The hybrid model outperforms the feature percolation and the grammaticality bias model only if we assume that the distortion rate is relatively high (>20%); see section Prior sensitivity analysis.

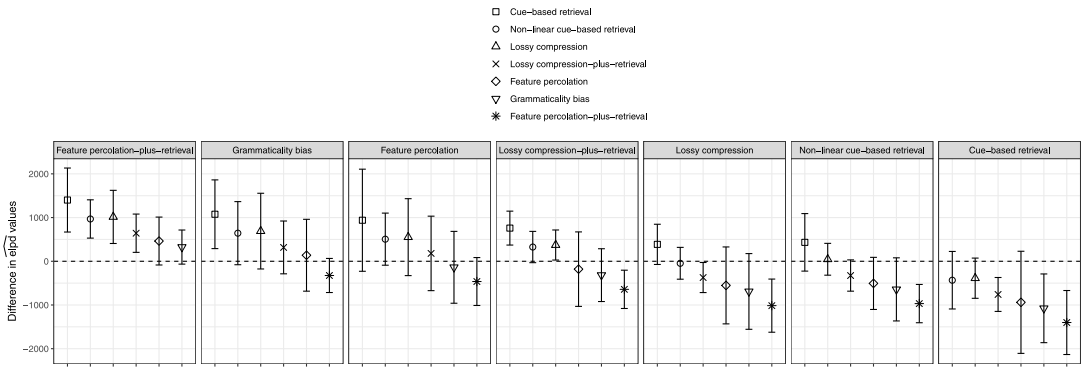


Fig. 13. The difference in predictive accuracies of the models represented by $\Delta\widehat{elpd}$ values: A positive $\Delta\widehat{elpd}$ value means that the model shown in the facet's title performs better than the other model. Error bars show two times the standard error of the difference in $\widehat{elpd}$ values.

A clear implication is that explanations for number agreement effects that exclusively invoke cue-based retrieval are ruled out: probabilistic representation distortion must be added to a theory of subject–verb number agreement processing. Previous modeling and empirical work has also suggested that representation distortion is necessary for explaining number agreement effects (e.g., Lago et al., 2021; Paape, Avetisyan, Lago, & Vasishth, 2021; Vasishth, Jäger, & Nicenboim, 2017; Villata, Tabor, & Franck, 2018).

Our work provides converging evidence broadly consistent with the general idea of representation distortion, but goes beyond this prior research in several important ways. First, ours is the first to use a computational implementation of multiple competing accounts to carry out a large-scale quantitative model comparison. Second, our model evaluation allows us to distinguish between two alternative accounts of representation distortion, lossy compression and feature percolation. Despite the fact that lossy compression is an important special case of representation distortion in sentence processing (Futrell et al., 2020), its predictive performance has, to our knowledge, never been evaluated in number agreement processing. Third, the numerically best-performing hybrid model is novel in that it combines feature percolation and cue-based retrieval. To our knowledge, no previous computational models of representation distortion in number agreement have combined feature percolation and cue-based retrieval to explain the patterns in the existing data.

The present work also advances our understanding of one well-known proposal in the literature on number agreement (Wagers et al., 2009). Under this view, retrieval occurs only in the ungrammatical sentence once the unexpected plural verb is encountered: the reader predicts a singular-marked auxiliary after reading *The key to the cabinets*. . . , and if the sentence continues with *are*, an error signal is raised, triggering a retrieval. This retrieval leads to misretrieval of the non-subject noun *cabinets* due to a number feature match of the verb with that noun. By contrast, in the grammatical sentence, no retrieval is triggered because once one has read *The key to the cabinets*. . . , encountering *is* leads to no error signal and therefore no retrieval: the singular verb that was expected is the one that is encountered. Thus, the Wagers et al. (2009) proposal relies only on cue-based retrieval and critically depends on the assumption that the features on the subject noun remain intact over time. This intact-feature assumption is untenable given our modeling results.

In fact, the Wagers et al. (2009) proposal was already difficult to reconcile with the broader literature on cue-based retrieval because, as also discussed in Villata and Franck (2020), there exists an array of data showing retrieval-driven processing difficulty in grammatical sentences (Mertzen, Paape, Dillon, Engbert, & Vasishth, 2022; Van Dyke, 2007; Van Dyke & McElree, 2006, 2011).

However, a reviewer suggests that these findings from Van Dyke and others cannot be invoked to challenge the Wagers et al. proposal. Recall that the key idea in Wagers et al.'s account is that in ungrammatical sentences, a particular number marking is predicted for the upcoming verb, and it is the mismatch between the input and the prediction that triggers the retrieval in ungrammatical conditions. The reviewer suggests that perhaps prediction does not occur in the Van Dyke design; if this were the case, then indeed, one cannot bring up the Van Dyke type design to challenge the Wagers et al. (2009) account: In the Van Dyke type grammatical sentences, if no prediction is made at all, retrieval would be carried out at the verb; this is in contrast to the grammatical number agreement design, e.g., in Wagers et al. (2009), where a number marking is predicted at the upcoming verb, and this prediction turns out to be correct, leading to no retrieval.

However, it seems very unlikely that no prediction occurs in the grammatical constructions used in the Van Dyke type design. Predictive processing is a well-established property of the human sentence comprehension system (e.g., Resnik, 1992). The central role that prediction plays in sentence comprehension has been demonstrated across a wide range of different syntactic configurations (see Hale, 2001; Levy, 2008a; Linzen & Jaeger, 2016; Staub, 2010a, among others). It would be odd to assume that the Van Dyke (2007) design differs from most other designs by somehow preventing prediction. Furthermore, as discussed in Mertzen, et al. (2022), there are good reasons to assume that prediction is happening in the Van Dyke designs: the effects of the interference manipulation show up consistently in the pre-critical region.[14] Under the cue-based retrieval account, one plausible reason why interference effects could show up in the pre-critical region would be that the verb features are already known (predicted) before the verb is encountered. This is because the retrieval process uses the feature specification of the verb to trigger a search for the target noun in memory. If the verb features are already available in the pre-critical region due to prediction, they could be used to retrieve the target noun, and consequently, the retrieval-based effects would show up in the pre-critical region.[15]

If we grant that prediction can occur in the Van Dyke designs, one could still argue that this prediction is not strong enough to avoid the retrieval process. Such an argument could follow from the Wagers et al. (2009) proposal: " .... When the verb is encountered its number features can be checked against the predicted features, and if they

[14] Also see the replications in Mertzen, et al. (2022) using the Van Dyke design in English and German.

[15] There could be other explanations as well for the effects observed in the pre-critical region, e.g., encoding interference (see the discussion in Mertzen, et al., 2022, p.34–39).

match, nothing more needs to be done; in particular, there is no need to retrieve material from the prior context. ...". If the Wagers et al. proposal assumes that the prediction is absolute, i.e., the verb features are always correctly predicted in grammatical cases, then no retrieval effect should occur in the Van Dyke designs. On the contrary, if the strength of prediction matters, one could argue that the prediction in the grammatical number-agreement design is much stronger than in the Van Dyke design; such an additional assumption would imply that retrieval does not occur in grammatical number agreement designs but does occur in the Van Dyke type design. But such an assumption raises a serious modeling problem: exactly how strong should be prediction in order to avoid triggering the retrieval (reanalysis) process? We think such a proposal (invoking strength of prediction) would be underspecified about how the initial agreement computation is driven by a predictive process. Therefore, under any of these assumptions, it is difficult to reconcile the Wagers et al. (2009) proposal with the data from Van Dyke and others showing that retrieval occurs in grammatical sentences.

Finally, there is a technical problem with the conclusion that no retrieval occurs in grammatical sentences: the statistical inference is based on null results from studies that are very likely underpowered (Jäger, Engelmann, & Vasishth, 2017).[16] Nevertheless, even though the Wagers et al. (2009) proposal in its original form is not tenable, it was the first sentence comprehension study to empirically demonstrate that cue-based retrieval could play a role in number agreement processing. Our work can be seen as building on and extending this important insight.

General discussion

What assumptions are necessary to explain the highly variable patterns of number agreement data observed in reading comprehension? To answer this question, we implemented two broad classes of models and proposed a new class of hybrid models. We compared the performance of these models on 17 published datasets to answer our question. The models and their key assumptions are summarized below:

1. Representation distortion-based models (Existing proposals)
    (a) Feature percolation model: The feature representation of the subject noun changes in memory due to probabilistic feature percolation from a non-subject noun in the pre-verbal input.
    (b) Grammaticality bias model: The continuous-valued representation of the subject's number changes due to the spread of plurality from a non-subject noun; an increase in number-mismatch between the subject and the verb causes an exponential increase in processing difficulty at the verb.
    (c) Lossy compression model: The feature representation of pre-verbal linguistic input can change probabilistically when it is stored in memory and the comprehender predicts the upcoming verb based on this lossy memory representation of the true pre-verbal context.
2. Cue-based retrieval models (Existing proposal)
    - Dependency completion between the subject and the verb is driven by a content-addressable search in memory based on feature specifications at the verb.
    - The feature representations on the noun remain intact over time.
3. Hybrid distortion-plus-retrieval models (New proposals)
    (1) Lossy compression-plus-retrieval model: The representation of preverbal linguistic input changes due to information loss before retrieval is triggered at the verb.
    (2) Feature percolation-plus-retrieval model: The representation of the subject noun changes probabilistically due to feature percolation before retrieval is triggered at the verb.

Table 3 shows the results of the model evaluation. A key result is that the hybrid feature percolation-plus-retrieval model shows the best performance numerically. Why does this hybrid model achieve the best fit? The reason is the model's behavior in grammatical sentences. In the sentences *the key to the cabinets was rusty* and *the key to the cabinet was rusty*, the retrieval mechanism of the model tries to retrieve the singular subject noun when the verb is encountered. In sentence *the key to the cabinets was...*, the probabilistic percolation of the plural feature sometimes changes the subject noun to plural, which causes difficulty in retrieval at the verb because there is no noun that fully matches the retrieval cues. In other trials, no feature percolation occurs, and the singular subject can be retrieved quickly. In sentences like *the key to the cabinet was...*, there is no plural feature that could percolate, so the verb is processed consistently slowly due to the fan effect from the second singular noun *the cabinet*. Consequently, the effect of distractor number in grammatical sentences is predicted to be neither decisively positive nor decisively negative; it fluctuates around 0. The data from 17 individual studies is consistent with this prediction.[17]

Fig. 14 shows the joint prior predictive distributions of this hybrid model for the agreement attraction effect (ungrammatical sentences) and the number distractor effect (grammatical sentences). The range of predicted effects covers the range of variation observed in the literature (compare to Figs. 1 and 11).

Another important result is that the grammaticality bias model outperforms the cue-based retrieval model of Lewis and Vasishth (2005), and the model's performance is statistically indistinguishable from the numerically best-performing hybrid model. This result implies that the grammaticality bias is one of the best candidates for explaining the observed number agreement effects. As we have demonstrated earlier in Fig. 7, the prior predictions of the grammaticality bias model are largely consistent with the qualitative pattern of effects in both grammatical and ungrammatical sentences.

Why does the grammaticality bias model achieve such a good fit to the data considered here? The model assumes that the processing cost at the verb increases exponentially as a function of the number mismatch between the subject and the verb. Consequently, a large number mismatch produces a large processing cost at the verb, but a small mismatch induces an almost negligible cost at the verb. This assumption predicts that the grammatical sentences *(a) the key to cabinet is rusty* and *(b) the key to cabinets is rusty* would not differ much in reading times at the verb; this is because the number mismatch between the subject *the key* and the verb *is* is quite small. Thus, even when the subject noun in (b) receives some plurality from *the cabinets*, it does not cause a big difference in processing cost at the verb compared to (a). However, in the case of ungrammatical sentences like *(c) the key to cabinets are rusty* vs. *(d) the key to cabinet are rusty*, the reading times at the verb would have a relatively large difference; this is because the number mismatch between the subject and the verb is high. As a consequence, compared to (d) a minor reduction in the number mismatch in (c) due to the spread of the plural feature from *the cabinets* would induce a large reduction in the processing cost at the verb. As

[16] In fact, a large-sample study using self-paced reading did find a 9 ms [0,18] effect of number agreement, consistent with the predictions of cue-based retrieval (Nicenboim, Vasishth, Engelmann, & Suckow, 2018).

[17] The non-linear cue-based retrieval model also predicts an effect fluctuating-around-zero in grammatical sentences but this model entirely fails to capture the agreement attraction effects in ungrammatical sentences (see Fig. 4).

Table 3
The differences in expected log predictive densities ($\widehat{elpd}$) of the models. The decisive evidence and the winning model are highlighted in bold face. The positive values imply evidence in favor of models written in the leftmost column compared to the models in topmost row. Also shown in brackets is the 95% confidence interval of the estimate of the difference in $\widehat{elpd}$.

	Cue-based retrieval	Non-linear cue-based retrieval	Lossy compression (LC)	LC+retrieval	Feature percolation (FP)	Grammaticality bias	FP+retrieval
Cue-based retrieval							
Non-linear cue-based retrieval	433 [−225, 1091]						
Lossy compression (LC)	387 [−73,847]	−46 [−410, 318]					
LC+retrieval	**759 [371, 1147]**	326 [−32, 684]	**372 [28,716]**				
Feature percolation (FP)	939 [−229,2107]	506 [−90, 1102]	552 [−328, 1432]	180 [−672, 1032]			
grammaticality bias	1076 [290, 1862]	643 [−79, 1365]	690 [−176, 1556]	317 [−287, 921]	138 [−684, 960]		
FP+retrieval	**1401 [669, 2133]**	**968 [530, 1406]**	**1015 [407, 1623]**	**642 [204, 1080]**	463 [−85,1011]	325 [−65, 715]	

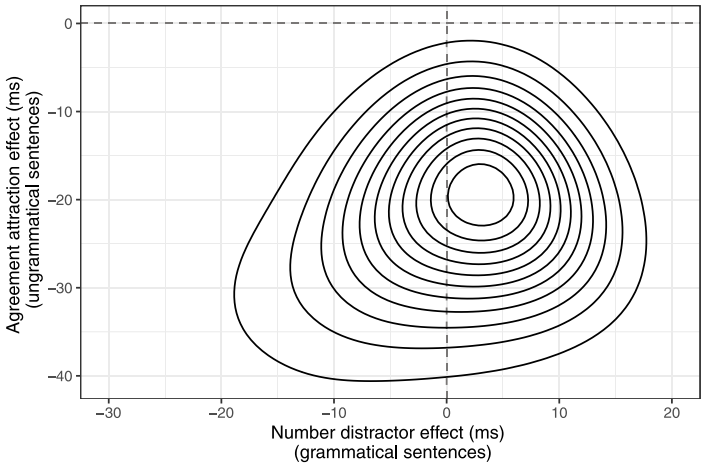


Fig. 14. The range of number agreement effects (in milliseconds) predicted by the feature percolation-plus-retrieval model is shown as a contour of the joint distribution of effects in the grammatical and ungrammatical conditions.

mentioned above, this happens because of the exponential nature of the processing cost function. Overall, the grammaticality bias model would predict an asymmetry between grammatical and ungrammatical sentences, which is consistent with the observed pattern of number agreement effects (see Fig. 7).

In sum, the quantitative model evaluation revealed two insights: First, cue-based retrieval alone (assuming intact representations) is insufficient for explaining the data. A representation distortion assumption is necessary: the subject noun's feature representation gets distorted through a process of percolation (feature migration). Second, adding a feature percolation assumption within the cue-based retrieval model makes it a good candidate explanation for the number agreement effects without losing the model's original empirical coverage of a variety of other constructions (e.g., see Vasishth & Engelmann, 2022). Given its overall empirical coverage, cue-based retrieval remains an important, independently motivated theoretical construct in models of sentence processing, although the grammaticality bias model is also competitive if one limits the comparison to the subject–verb agreement constructions considered here. We discuss these points next.

The representation distortion assumption is necessary in dependency completion theories

A priori, it seems reasonable to question the assumption that memory representations get distorted over time. Is such a position justifiable when we go beyond number agreement and look more broadly at the literature on sentence processing? Is there any other evidence in sentence processing, and more generally in working memory research within cognitive psychology, that input strings get probabilistically distorted?

In the early days, sentence processing theories generally took a different direction. From the 1970s onwards, a default (if implicit) assumption was that feature representations of nouns remain intact, and a single veridical representation is the end-result when one parses a sentence (e.g., Frazier, 1979). Early models like the garden-path theory (Frazier, 1987) implicitly assumed a deterministic parsing process that always carried out the same steps given a particular input string. The same holds for the well-known Dependency Locality Theory (Gibson, 2000): under this account, no aspect of the input string ever experiences any distortion; the only metric that quantifies dependency completion difficulty is memory load (storage cost and/or integration cost), computed deterministically.

A significant departure from this intact-representation perspective gained momentum with the arrival of alternative accounts like good-enough processing (Ferreira, Ferraro, & Bailey, 2002). Good-enough processing makes a major departure from this classical view by assuming that once non-veridical syntactic representations of the input string are created in memory, these mis-parses remain in memory as retrievable chunks. For example, when we read *While Mary bathed the baby played in the crib*, an initial mis-parse assigns *the baby* as an object to *bathed*, but a subsequent reanalysis undoes this mis-parse and *the baby* becomes the subject of the main verb *played*. Interestingly, if the participant is asked whether Mary bathed the baby, the participant tends to give the incorrect answer "yes". This suggests that non-veridical representations of an initial mis-parse can remain in memory. As Ferreira et al. (2002) put it: "the meaning people obtain for a sentence is often not a reflection of its true content".

More recent models of comprehension that rely on Bayesian inference (so-called rational inference, Anderson, 1991), such as the noisy channel model (Gibson, Bergen, & Piantadosi, 2013; Levy, 2008b) and

the lossy-context surprisal model (Futrell et al., 2020), go even further: the input string can be modified by inserting or deleting material, changing the string into one that is informed by their prior beliefs about what is likely to occur in their language. For example, if one reads *The dog was bitten by the man*, one might mentally reverse the roles of the biter and bitee to match real-world probabilities of events and perceive the sentence as meaning *The man was bitten by the dog*. There are also connectionist models that assume that the relative plausibility of event representations drive the semantic interpretation of a sentence (Rabovsky, Hansen, & McClelland, 2018).

Thus, in recent decades, representation distortion at the sentence level has become an increasingly plausible explanatory construct for theories of comprehension. These accounts – good-enough processing, noisy channel, lossy compression, etc. – generally focus on the representation distortion of the entire input string, not just the internal feature representations of individual words in the string.

In contrast to assumptions about sentence-level distortion of the input, research in the cognitive psychology of memory has historically focused on understanding the constraints on single isolated units of information, such as words, letters, numbers, visual shapes, etc. Jonides, et al. (a comprehensive review is in 2008). Even when memory researchers look beyond these minimal units of information, they usually investigate extremely simple syntactic frames like *A is B* (Anderson et al., 1983). Despite this narrow focus, theories of memory in cognitive psychology have evolved into quite a sophisticated account of how these units of information experience representation distortion. For example, the feature overwriting model (Nairne, 1990) formalizes the idea of feature-level representation distortion (also see Oberauer & Kliegl, 2006). As Nairne (1990, p. 252) puts it: *An individual feature of a primary memory trace is assumed to be overwritten, with probability F, if that feature is matched in a subsequently occurring event. Interference occurs on a feature-by-feature basis, so that, if feature b matches feature a, the latter will be lost with probability F*. This implies that the more similar two items are, the harder they are to distinguish, which will make both harder to retrieve given some retrieval cues.

There is also independent evidence in memory research that supports some form of feature transfer from one item to another in memory. Researchers in psychology have investigated what kind of errors occur when participants have to report the feature(s) of target item that was recently presented along with a distractor item. An important finding is that in some proportion of trials, participants make *swap errors*: they mistakenly report the features of the distractor item when probed about the target item (Bays, 2016; Bays, Catalao, & Husain, 2009; Scotti, Hong, Golomb, & Leber, 2021). Swap errors support the idea that feature migration among items may explain why representations get distorted in working memory.

Within psycholinguistics, it was only some 30 years ago that the connection between memory research in psychology and the constraints on sentence processing was articulated (Lewis, 1993, 1996), although the focus was primarily on interference effects at the sentence level. The importance of word-level feature encoding in sentence processing gained prominence through a thread of research that falls under the broad rubric of *encoding interference* (e.g., Barker, Nicol, & Garrett, 2001; Gordon, Hendrick, & Johnson, 2001; Hofmeister & Vasishth, 2014; Jäger, Benz, Roeser, Dillon, & Vasishth, 2015; Smith et al., 2021; Villata et al., 2018). The idea – which derives from feature overwriting and other related accounts (e.g., Nairne, 1990) – is that if two nouns have similar features (such as two animate nouns), they will be more difficult to maintain in memory compared to the case where the nouns have no overlapping features (e.g., animate vs. inanimate nouns).

This general idea of encoding interference appeared in number agreement research as well, but the initial theoretical explanations came from sentence production; memory processes were considered, but only cursorily (Bock & Miller, 1991). In production, the puzzle that needed an explanation was that people tend to write or utter ungrammatical constructions like *The key to the cabinets are rusty*. As discussed earlier, this agreement attraction phenomenon eventually gained importance in empirically driven sentence comprehension theory as well (e.g., Clifton, Frazier, & Deevy, 1999; Nicol, Forster, & Veres, 1997; Pearlmutter et al., 1999), and representation distortion accounts like feature percolation (Eberhard, 1997; Franck, Vigliocco, & Nicol, 2002b; Nicol, 1995; Nicol et al., 1997) and the closely related Marking and Morphing model of Eberhard et al. (2005) (see also Bock, Eberhard, & Cutting, 2004; Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001; Brehm & Bock, 2013; Eberhard et al., 2005; Staub, 2009) became candidate explanations in work on comprehension. In the agreement literature, the cue-based retrieval account was invoked much later (Wagers et al., 2009). The close connection between the representation distortion accounts in number agreement and the existing theories of feature overwriting within cognitive psychology only became apparent in subsequent work (e.g., Smith et al., 2021; Villata et al., 2018).

In summary, it seems that, in language processing, as in memory research, there are in fact good reasons to assume that some form of probabilistic representation distortion occurs at both the word and sentence level. Considering that there is plenty of independent evidence for representation distortion from pure memory research in cognitive psychology, it therefore seems reasonable that some type of feature overwriting mechanism should be an integral part of sentence comprehension models. More focused experimental work is required to understand the underlying representation distortion process in sentence comprehension, e.g., what types of distortion are possible and under what circumstances distortion can occur.

Cue-based retrieval is an important component of sentence processing

A key finding of the present work is that, when combined with representation distortion, cue-based retrieval becomes one of the best candidates for explaining number agreement effects. With that finding and the discussion of encoding interference above, one interesting observation here (articulated in Laurinavichyute, 2021; Lewis et al., 2006; Villata et al., 2018) is that encoding and retrieval accounts are not opposing but complementary explanations, and both processes could be in operation simultaneously. As a consequence, it seems obvious that the existing cue-based retrieval architecture should also have a representation distortion mechanism that probabilistically distorts a noun's feature specification. The best-performing models across prior specifications always include some mechanism for feature distortion in addition to cue-based retrieval, which, on its own, was not enough to explain the full pattern of reading times. What was missing until now was a comprehensive demonstration that the additional complexity of representation distortion is needed in the model. Providing such a demonstration is one of the achievements of the present paper.

Our hybrid feature-percolation-plus-retrieval model is a minimal modification of the original cue-based retrieval model and adds only one parameter to the model. The hybrid model assumes that the subject noun's number feature is probabilistically distorted according to a distortion rate parameter. In any particular trial, regardless of whether distortion occurs or not, the standard cue-based retrieval constraints will apply, and reading time will be determined by the classical constraints on retrieval derived from the cognitive architecture ACT-R. Embedding the representation distortion assumption into the cue-based retrieval model has a great advantage because the cue-based retrieval model can already account for a broad range of empirical data from a variety of constructions, including subject–verb non-agreement dependencies (Mertzen, et al., 2022; Van Dyke, 2007; Van Dyke & McElree, 2011), plausibility mismatch configurations (Cunnings & Sturt, 2018), negative polarity item licensing (Drenhaus, Saddy, & Frisch, 2005; Vasishth, Brüssow, Lewis, & Drenhaus, 2008; Xiang, Dillon, & Phillips, 2009), and honorific processing (Kwon & Sturt, 2016), in both unimpaired and impaired populations such as individuals with aphasia (Vasishth et al., 2019). Assuming that only encoding

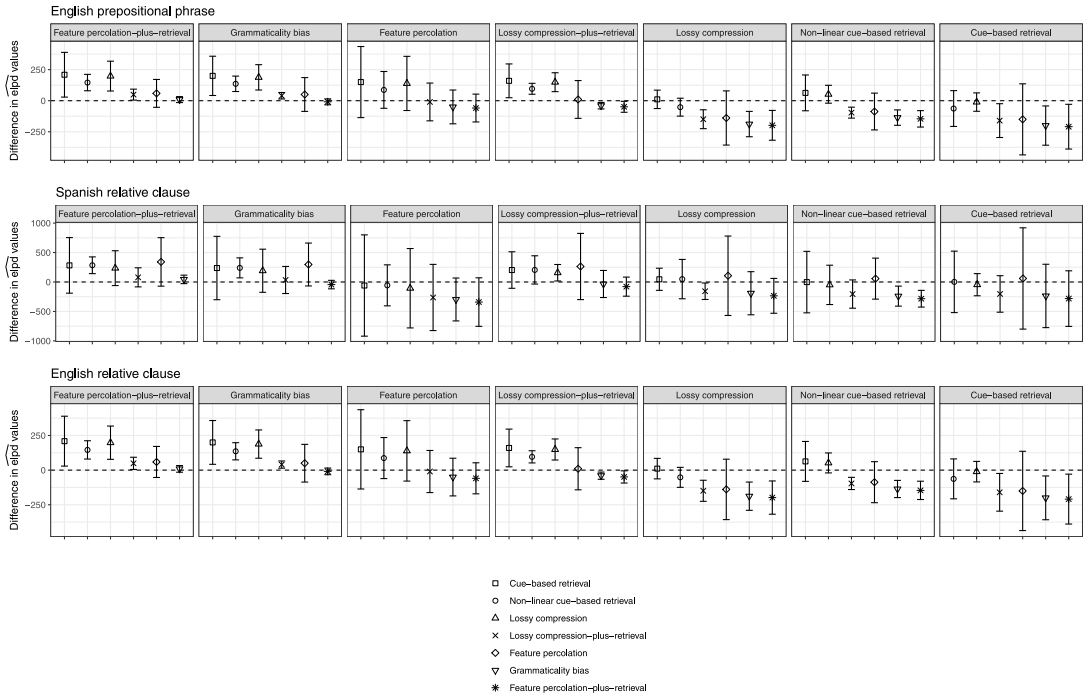


Fig. 15. The difference in predictive accuracies of the models represented by $\widehat{\Delta elpd}$ values for each design: A positive $\widehat{\Delta elpd}$ value means that the model shown in the facet's title performs better than the other model. Error bars show two times the standard error of the difference in $\widehat{elpd}$ values.

interference is in operation would lead to poorer empirical coverage than the hybrid architecture.

However, the comparable performance of the hybrid model and the distortion-based models (see Fig. 13) indicates that number agreement effects can be potentially explained by a distortion-based mechanism alone, without requiring the cue-based retrieval assumption. One might conclude that cue-based retrieval theory can be abandoned completely as a theory of dependency completion. However, as discussed above, cue-based retrieval is a much more general theory of dependency completion than the grammaticality bias model, which seeks to explain only one phenomenon: number agreement.

Even for the number agreement dependency considered here, a minor modification to the cue-based retrieval process, i.e., adding a feature distortion mechanism in the model, makes it the numerically best-performing model given the data. This hybrid distortion-plus-retrieval model also captures the qualitative pattern of number agreement effects in both grammatical and ungrammatical sentences (see Fig. 11). Therefore, if we want as general a theory of number agreement as possible, cue-based retrieval remains an important component of the explanation.

The distortion rate may differ across designs and across participants

An interesting open question is whether the rate of feature distortion differs across experimental designs and among individual participants in a study. There are good reasons to think that the percolation rate might differ across experimental designs because it is arguably sensitive to factors like syntactic distance between the nouns and syntactic position of the nouns relative to each other (Franck, Lassi, Frauenfelder, & Rizzi, 2006; Franck, Soare, Frauenfelder, & Rizzi, 2010; Franck, Vigliocco, & Nicol, 2002a). It has been suggested that there are constraints on how far a number feature can migrate in a tree (Eberhard et al., 2005; Nicol et al., 1997). For instance, a feature from a noun inside a prepositional phrase can migrate more easily compared to that from a noun in an embedded relative clause. Another proposal is that a number feature is constrained to migrate only upwards in a tree (Eberhard, 1997; Vigliocco, Butterworth, & Semenza, 1995). Similar constraints can also be implicated for the deletion rate and the insertion rate parameters in the lossy compression models.

Our models are currently agnostic to these factors. More specifically, we make a *cross-design homogeneity assumption*: the data for all 17 studies, regardless of their construction type and language, are assumed to come from the same underlying (true) model and same parameter value(s). This assumption may lead to inaccurate generalization if there is any systematic across-design variation in the underlying process. Would our model comparison results still hold if we do not assume the cross-design homogeneity?

We can test this concern using a within-design model comparison as suggested by one of the reviewers. We consider three designs which were employed in at least three experiments: (a) English prepositional phrase, (b) English relative clause, and (c) Spanish relative clause. Within each design, the models are evaluated in terms of their predictive performance: Each model is fitted and tested on the design-specific data using cross-validation. So that even if there is any cross-design parametric variation, it does not impact the model evaluation performed within a design.

Fig. 15 shows the model comparison results within each of the three designs considered here. We find that the models' predictive performance patterns are similar to the results obtained under the cross-design homogeneity assumption (Figure 13): the models assuming some kind of feature distortion outperform the cue-based retrieval models, and the hybrid feature percolation-plus-retrieval model and the grammaticality bias model are the best-performing models numerically. The analysis indicates that our main results hold independent of the cross-design homogeneity assumption.

A noteworthy result from the within-design analysis is that the grammaticality bias model performs almost as well as the hybrid feature percolation-plus-retrieval model, which is numerically the best-performing model. This parallels the results for the cross-design homogeneity analyses. Together, the results raise the question of which mechanism best explains the number agreement effects, the hybrid feature distortion-plus-retrieval, or a grammaticality bias mechanism? We cannot conclusively answer this question given the current results. As noted above, cue-based retrieval is an empirically-grounded theory of dependency completion with support from a variety of sentence processing effects. Thus, the hybrid mechanism involving a retrieval process would have better generalizability across different construction types. On the other hand, grammaticality bias is a fresh, emerging perspective on the agreement attraction phenomenon (see Hammerly et al., 2019). The proposal holds promise to become a new theory of dependency completion, where the cost of dependency completion is a non-linear function of the quantitative feature mismatch between the two co-dependents. However, for the grammaticality bias model to become a theory of dependency completion in general, a detailed empirical investigation will be necessary. This empirical investigation would have to consider model fit to data on other kinds of dependencies, such as subject–verb non-agreement dependencies (Mertzen, et al., 2022; Van Dyke, 2007; Van Dyke & McElree, 2011), antecedent-reflexive dependencies (Dillon et al., 2013; Sturt, 2003), plausibility mismatch configurations (Cunnings & Sturt, 2018), and negative polarity constructions (Vasishth et al., 2008).

The rate of feature distortion can also differ across individuals: some individuals can be better at retaining the original representation of nouns in memory than others. Systematic modeling of individual differences can reveal important insights about the underlying process (Yadav, Paape, Smith, Dillon, & Vasishth, 2022). For example, one can model individual differences in number agreement effects as variation in the percolation rate parameter of the feature percolation-plus-retrieval model. We plan to investigate design-level and individual-level differences in feature migration in future work.

The markedness effect in number agreement processing

A well-explored issue in the number attraction literature is the *markedness effect*: the overtly marked nouns, e.g., plural nouns can cause agreement attraction but not the unmarked nouns (e.g., singular nouns) (Bock & Eberhard, 1993; Eberhard, 1997; Wagers et al., 2009, inter alia). The effect has been frequently observed in production studies on agreement attraction (Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991). However, in the reading studies considered here, there is only one experiment that demonstrated the markedness effect, experiment 3 in Wagers et al. (2009) (but also see Pearlmutter et al., 1999). A prominent explanation for the markedness effect is that only the overtly marked features can migrate from the attractor noun to the subject noun (Eberhard, 1997). For example, when the attractor noun is singular, the number feature is unmarked, hence the singular feature cannot percolate to the subject noun. The feature percolation model would predict no attraction effect in this situation. Similarly, the feature-percolation-plus-retrieval (FPR) model would predict the same attraction effects as the cue-based retrieval model because the percolation rate would become zero. Therefore, the feature percolation-based models predict an asymmetry in the attraction effects in design A vs. design B (see Fig. 16) under the markedness assumption.

In this work, we do not make an explicit markedness assumption in our models. This is because we want our models to be simple, generalizable, and have fewer constraints so that the models can be easily extended to generate predictions for interference in other construction types, such as non-agreement subject–verb dependencies (Van Dyke, 2007; Van Dyke & McElree, 2011), dependencies involving a semantic plausibility manipulation (Cunnings & Sturt, 2018), and antecedent-reflexive dependencies (Cunnings & Felser, 2013; Dillon et al., 2013; Jäger et al., 2020).

However, if needed one can easily build the markedness constraint in our models by simply assuming that the feature percolation rate is zero whenever the distractor/attractor is singular. Does the markedness assumption provide a better fit to the data? To test this, we implement the feature percolation and the FPR model under the markedness constraint and then compare them against the corresponding unconstrained models. This approach would allow us to infer whether the markedness assumption is better at capturing the observed data. We find that the models-with-markedness perform only slightly better than their unconstrained counterparts: the $\widehat{elpd}$ value for the feature percolation-with-markedness model vs. the feature percolation model is $71[-37, 179]$ and for the FPR-with-markedness vs. the FPR model is $28[-16, 72]$. Thus, given the available data, there seems to be little reason to add the markedness assumption in our feature percolation-based models.

Can the lossy compression assumption explain number agreement effects?

In our model evaluation, the lossy compression model performs better than the cue-based retrieval model. What makes this model achieve a better fit to the number agreement data?

Consider the grammatical sentences *(a) the key to the cabinet was rusty* vs. *(b) the key to the cabinets was rusty*. The model's behavior can be understood as an interaction between representation distortion and the probabilistic expectation at the verb. When there is no information loss and the representation of the nouns remains intact, the model relies on only the probabilistic expectation of the upcoming verb. Based on corpus statistics, the singular verb is more expected to occur after pre-verbal contexts like *the key to the cabinet*, where both nouns are singular. Consequently, when the distortion rate is small, the model predicts a higher processing difficulty in sentence (b) compared to (a) because processing difficulty is inversely proportional to how expected the verb is given the preceding context. As the distortion rate increases, this effect starts diminishing (see Fig. 17, right panel). For example, as the rate of inserting a plural marker increases, the pre-verbal input in (a) *the key to the cabinet* is more likely to get distorted to other possible contexts that generate less accurate predictions and hence cause more processing difficulty. The model therefore predicts an effect consistent with negative, zero, or a small, positive effect in grammatical sentences; this prediction captures the data better than the cue-based retrieval model, which only predicts negative effects.

However, the lossy compression model performs worse than the hybrid distortion-plus-retrieval models. A follow-up study is needed to explore whether some other assumptions about the nature of information loss can improve the model's fit; for example, whether the representation gets distorted due to only insertion noise or due to both insertion and deletion noise. In our implementation of lossy compression, we observe that the insertion rate parameter modulates the magnitude of number agreement effects; see Fig. 17. As the insertion rate increases, the number distractor effect (in grammatical sentences) decreases and the agreement attraction effect (in ungrammatical sentences) increases. In contrast, the deletion rate parameter does not seem to influence number agreement effects, indicating that only insertion noise might be driving the effects in subject–verb number agreement.

Another assumption worth exploring is that the information loss may increase over time. For example, it is possible that the representations of nouns that appear early in sentence are more likely to get distorted compared to the nouns that appear later. This could happen because the nouns that appear earlier have been stored in memory for a longer time and thus experience more information loss than the nouns that appear later (Futrell et al., 2020). We plan to implement and compare different versions of the lossy compression assumption on agreement data in a future study.

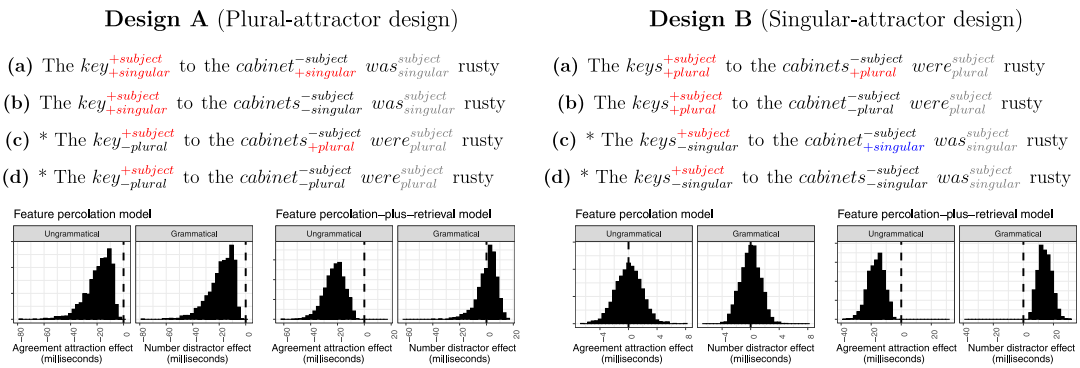


Fig. 16. Prior predictions of the feature percolation and the hybrid model under the markedness constraint: When the percolation-based models assume *markedness*, they predict a different pattern of effects in Design A vs. Design B; in Design B, the singular feature of the non-subject noun in conditions (b) and (c) cannot percolate to the subject.

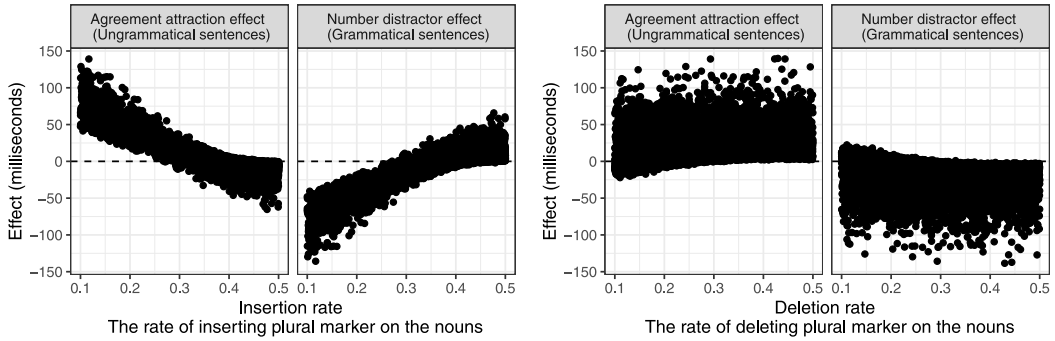


Fig. 17. The agreement attraction and the number distractor effect predicted by the lossy compression model as a function of insertion rate (right) and deletion rate (left).

Desiderata for a complete theory of dependency completion

The work reported here is a detailed investigation of one type of syntactic dependency, involving subject–verb number agreement. The theoretical ideas developed here should ideally have broader application across sentence processing more generally. What should a more general theory look like? What should it be able account for? We discuss this question next.

A complete theory of dependency completion in sentence comprehension should have at least the following three properties. First, it should capture the joint distribution of effects in grammatical and ungrammatical sentences. Second, it should be able to account for systematic variation in individual-level effects. Third, it should explain data across different constructions, such as argument-verb dependencies, antecedent-pronoun and antecedent-reflexive dependencies, etc.

We have focused on the first property in this work. We asked which theoretical assumptions find the strongest support in the data from both grammatical and ungrammatical sentences. Modeling this joint distribution is critical for a theory of subject–verb number agreement processing because it holds some crucial information about the underlying comprehension process. As we can see in Fig. 1, the effect in ungrammatical conditions tends to be facilitatory but the effect in grammatical conditions fluctuates around zero. Although there is a tendency in the literature to oversimplify the grammatical results as showing that the effect is 0 ms (Hammerly et al., 2019), this simplified conclusion ignores the uncertainty of the estimates in the 17 studies, and their variability. A theory that can account for the range of variation observed (in both grammatical and ungrammatical conditions) may be a more realistic account of the underlying cognitive processes in sentence comprehension.

The second property that a theory should explain (individual differences in observed effects) has historically been largely neglected in sentence comprehension. Modeling individual-level behavior can lead to new theoretical insights (Kidd, Donnelly, & Christiansen, 2018), and focusing only on average behavior can lead to inaccurate generalizations (Fific, 2014; Tanner, 2019). For example, in a recent study, Yadav, et al. (2022) have shown that a hypothesis inferred from averaged data – that syntactic cues are preferred over non-syntactic cues in processing antecedent-reflexive dependencies – holds only for some, not all, of the participants. In future work, one could experimentally obtain an independent measure of the feature percolation-plus-retrieval model's percolation rate parameter for each participant in a study using, for example, a participant's error rates on comprehension questions that directly probe the number marking on the nouns (Avetisyan et al., 2020). The individual-level percolation rate parameters estimated in this way could be used to generate reading time predictions from the model which could then be evaluated using the same participants' reading time data.

Finally, a complete theory of dependency completion should be generalizable to constructions other than subject–verb number agreement dependencies. An example is gender agreement dependencies. The current work has focused on subject–verb number agreement dependency because this construction has a relatively large amount of published data that is also publicly available.[18] Other dependencies have also been investigated for interference effects in comprehension, including subject–verb constructions in which syntactic and semantic similarity

[18] Data availability is important because, as documented in previous meta-analysis attempts that we have carried out (Jäger et al., 2017; Nicenboim, Roettger, & Vasishth, 2018), it is often difficult or even impossible to infer the relevant statistics from published analyses alone. Published results often focus on reporting statistical significance, and neglect to report mean differences and the standard errors of these differences.

is manipulated (Mertzen, et al., 2022; Van Dyke, 2007; Van Dyke & McElree, 2011); negative polarity items (Vasishth et al., 2008); and antecedent-reflexive dependencies (Cunnings & Felser, 2013; Dillon et al., 2013; Jäger et al., 2020).

A limitation of our work, and of research on number agreement more generally, is that the two feature percolation models have only been instantiated in terms of the percolation of a number feature. The feature percolation-based models cannot predict anything for constructions showing interference due to gender or semantic features (e.g., Cunnings & Sturt, 2018). However, one could relax this assumption of the model so that it is possible to percolate non-number features as well. In future work, we plan to extend the current models to evaluate them on data from non-agreement subject–verb and antecedent-reflexive dependencies. We turn now to other future directions before we conclude.

Future directions

Our model evaluation results imply two obvious investigations that must be carried out. First, how well do the competing models do against a much broader range of benchmark data that go beyond number agreement? Second, are there other models that can outperform the hybrid model presented above? These two future directions are not without their own challenges, as we discuss now.

Future model comparisons must deploy a broader spectrum of high-powered, open access benchmark data

One major barrier to testing model performance against a broader spectrum of benchmark data is that psycholinguistics, like other adjacent areas in cognitive science (Open Science Collaboration et al., 2015), is still in the process of catching up with the open access and transparency revolution that is unfolding in other areas of science. In addition, the psycholinguistic data that happen to be publicly available are usually severely underpowered (this is discussed in Jäger et al., 2017, 2020; Vasishth & Gelman, 2021; Vasishth, Mertzen, Jäger, & Gelman, 2018; Vasishth, Yadav, Schad, & Nicenboim, 2022). Low power, coupled with publication bias (Francis, 2012), leads to effect size estimates that are too large and are unlikely to reflect a realistic range of effect sizes, a phenomenon referred to as Type-M and Type-S error (Gelman & Carlin, 2014). Fortunately, more and more researchers are responding to this problem by running larger-sample experiments and carrying out direct replication attempts (e.g., Lago et al., 2021; Villata et al., 2018). Such higher-powered studies will be very useful as benchmark data for future model development.

New candidate models should be pitted against existing ones using quantitative methods

Several new computational models have emerged in recent years that attempt to explain number agreement data. Two prominent examples are self-organized parsing (Smith, Franck, & Tabor, 2018; Smith et al., 2021) and neural network models (Linzen & Dupoux, 2016; Ryu & Lewis, 2021). Current evaluations of these models are generally very limited in scope; the focus of the model evaluation is often restricted to modeling average error rates in production (Linzen & Dupoux, 2016) or qualitative patterns in reading times without regard to the magnitude and range of variation in the data (Ryu & Lewis, 2021; Smith et al., 2018, 2021). Current implementations of self-organized models of number agreement (Smith et al., 2018, 2021) make reading time predictions similar to those of the feature percolation model: a distractor noun with the same number marking as the subject always leads to slower processing than a distractor noun that differs in number from the subject. The implemented models assume a veridical representation of the input, although Smith et al. (2021), Villata and Franck (2020), Villata et al. (2018) discuss possible extensions that involve a form of feature distortion on the nouns in the subject NP. However, these ideas have not yet been implemented, so their exact predictions and predictive performance have not yet been determined.

The methodologies we have used in this work, approximate Bayesian computation for parameter fitting and cross-validation for model comparison, can easily be extended to other classes of models, facilitating future work comparing new classes of competing theory.

Conclusion

We have proposed a new model of agreement attraction that exhibits a superior fit to the number agreement data compared to existing competing models. We compared the predictive performance of seven models, including two new proposals, using benchmark data from 17 experiments that investigated subject–verb number agreement dependencies. The model comparison revealed two major theoretical insights. First, a well-accepted explanation of agreement attraction – cue-based retrieval – alone is insufficient for explaining the data, and an assumption that the nouns stored in memory can undergo feature distortion seems necessary. Second, number agreement effects are possibly caused by either (i) a hybrid mechanism such that the cue-based retrieval operates on the potentially-distorted representation of nouns in memory due to feature percolation, or (ii) a grammaticality bias during reading such that the cost of dependency completion increases exponentially as the feature mismatch between the two co-dependents increases. To our knowledge, this work is the first attempt at quantitatively evaluating the competing theories of number agreement processing using data from multiple studies simultaneously. The approach presented here can be easily adapted to compare computational models of any underlying cognitive process without compromising the complexity of the models.

CRediT authorship contribution statement

Himanshu Yadav: Conceptualization, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Garrett Smith:** Validation, Supervision, Writing – original draft, Writing – review & editing. **Sebastian Reich:** Formal analysis, Methodology, Writing – original draft. **Shravan Vasishth:** Supervision, Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code and data associated with this paper are available from https://osf.io/gqj3p/.

Acknowledgments

Our heartfelt thanks to Colin Phillips, Brian W. Dillon, Sol Lago, Matthew Tucker, Matthew Wagers, Nicole Patson, and Matthew Husband, who provided the raw data for this modeling work. We thank Scott Sisson for his important feedback on the methodology used in this work. We also thank Bruno Nicenboim and Dario Paape for their insightful comments on the initial draft. HY received funding from the Deutscher Akademischer Austauschdienst - DAAD, Germany, Program ID: 57440921, Reference no.: 91730718. SV and GS received funding from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany, Project-ID 317633480, SFB 1287. SR was partially funded by Deutsche Forschungsgemeinschaft (DFG), Germany, Project-ID 318763901, SFB 1294.

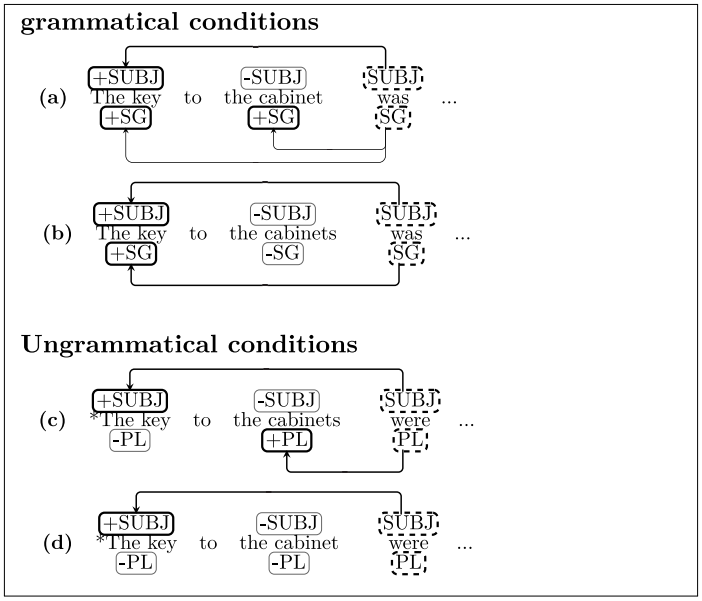


Fig. A.1. A schematic illustration of activation received by the nouns through each retrieval cue in the cue-based retrieval model of Lewis and Vasishth (2005). The dashed box represents the retrieval cues, the thick box represents the noun feature that matches the retrieval cue, the thin box represents the noun feature that does not match the retrieval cue. The thickness of the arcs from a retrieval cue to the nouns represent the amount of activation received by the nouns from each cue; the thicker lines mean larger activation received by the noun and vice versa. For example, in condition (a), both the nouns match the number cue, SG, and hence the activation gets divided among the nouns represented by thin lines. While in condition (b), the subject noun receives all the activation from the number cue which is represent by the thicker line.

The cue-based retrieval model

The cue-based retrieval model developed by Lewis and Vasishth (2005) adopts general principles of ACT-R cognitive architecture (Anderson, et al., 2004; Anderson & Lebiere, 2014). The model assumes that the dependency completion is driven by a cue-based retrieval process: a content-addressable search in memory based on feature specifications such as [subject], [plural], called retrieval cues. Each chunk in memory that matches a retrieval cue receives a certain amount of activation. The total activation of a memory chunk i is given by,

$$A_i = B_i + \sum_{j=1}^{N} W_j S_{ji} + \epsilon_i \tag{A.1}$$

where B_i is the base-level activation of the chunk; $W_j S_{ij}$ is the amount of activation received by chunk i from the retrieval cue j; $\sum_{j=1}^{N} W_j S_{ji}$ is the sum of activations received by the chunk through each retrieval cue; ϵ_i is the trial-level noise in the activation such that $\epsilon_i \sim Normal(0, \sigma)$.

The model further assumes that a chunk with the highest total activation gets retrieved and the activation of the retrieved chunk determines the retrieval time at the verb.

$$RT_i = Fe^{-A_i} \tag{A.2}$$

where A_i is the activation of the retrieved chunk i and F is a scaling parameter, called the latency factor. The latency factor is the only free parameter in the model in the present paper.

Using the above retrieval time equation, we derive predictions for number agreement effects in grammatical and ungrammatical conditions for subject–verb agreement dependencies. Suppose, the effect in grammatical and ungrammatical conditions is, δ_g and δ_u respectively. The effects come from the model conditioned on its one free parameter, the latency factor,

$$\begin{pmatrix} \delta_g \\ \delta_u \end{pmatrix} \sim Model(F) \tag{A.3}$$

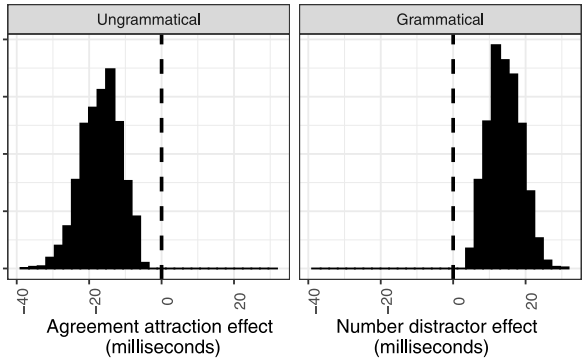


Fig. A.2. The agreement attraction and the number distractor effect predicted by the cue-based retrieval model.

where F – the scaling parameter – is assumed come from the following prior distribution,

$$F \sim Normal_{lb=0.05}(0.15, 0.05) \tag{A.4}$$

where $Normal_{lb=0.05}(0.15, 0.05)$ means a normal distribution with mean 0.15, standard deviation 0.05, and truncated at 0.05 so that it does not allow values lower than 0.05. The justification for the choice of priors on the scaling parameter is given in section Choice of priors on the scaling parameter .

Fig. A.1 shows the activation profiles for the grammatical and ungrammatical subject–verb number agreement dependencies. In condition (a), both the noun phrases, *the key* and *the cabinet* receive activation from the number cue [SG]. As a result, the amount of total activation available through the number cue gets divided among the two noun phrases. This is called the *fan effect* (Anderson, et al., 2004; Schneider & Anderson, 2012). Due to the fan effect, the subject noun in condition (a) receives less activation compared to condition (b),

where only the subject noun matches the number cue. Consequently, the retrieval at the verb is predicted to be slower in condition (a) compared to condition (b). Hence, the model predicts an inhibition due to the distractor noun in the grammatical conditions (see Fig. A.2).

In the ungrammatical sentences, in condition (c), the subject noun receives activation only through the subject cue [SUBJ] and the distractor noun phrase *the cabinets* receives activation only through the number cue [PL]. Hence, the two noun phrases receive an equal amount of activation, which leads to a race for retrieval between the two (Engelmann et al., 2019; Logačev & Vasishth, 2016). This race process produces faster retrieval times in condition (c) compared to (d) where there is no race for retrieval. So, the model predicts a facilitation due to the attractor noun in ungrammatical conditions (see Fig. A.2).

Choice of priors on the scaling parameter

All the models that we evaluated in this work have a free parameter called the scaling parameter. The scaling parameter maps the reading time output from the model on the same scale as the reading time distribution from experimental studies on subject–verb number agreement. How should we choose the priors on the scaling parameter? Our choice of priors on the scaling parameter for a model is based on the following two criteria:

1. **The model should not generate unreasonably fast reading times.** A typical distribution of reading times from the self-paced reading or eye-tracking experiments (e.g., total fixation time) has the property that the reading time values are usually larger than approximately 100–150 ms (chapter 6, Nicenboim et al., 2022). A model that generates a large proportion of reading times smaller than 100–150 ms would be an inaccurate characterization of the underlying generative process. Indeed, the reading times from agreement attraction studies have a 2.5th percentile of approximately 150 ms, and the 97.5th percentile varies between studies from 250 to 800 ms (see Fig. B.1). Following the prior distribution guidelines in Schad et al. (2021), we choose a prior on the scaling parameter such that the 2.5th percentile is approximately equal to 150 ms.
2. **All the models should generate approximately the same range of reading time distribution.** We want the models to generate reading times in a similar 95% credible interval. Having fixed the lower bound of the interval at 150 ms (see the previous point), we choose a reasonable credible interval of $[150, 300]$ ms; each model should generate reading times with a 95% credible interval of approximately $[150, 300]$ ms. We choose this range because we do not want our models to be wildly different in terms of their reading time distribution. This is necessary for calibrating models with the experimental reading times data and for drawing meaningful comparisons between their performance.

The non-linear cue-based retrieval model

We implement the cue-based retrieval model assuming direct-access and non-linear cue-combination proposed by Wagers et al. (2009) (also see Wagers, 2008). Wagers and colleagues speculate that a direct-access mechanism (McElree, 2000) where cues are combined non-linearly can capture the observed pattern of number agreement effects in grammatical and ungrammatical sentences. In order to implement their proposal, we use the earlier implementations of the direct access model (e.g., Lissón, et al., 2021; Nicenboim & Vasishth, 2018) and make some additional assumptions for non-linear cue-combination and processing in ungrammatical sentences.

The main assumption of the model is that subject–verb dependency completion is driven by a content-addressable search in memory based on feature specifications such as [subject], [plural], called retrieval cues. Each chunk in memory that matches a retrieval cue competes for retrieval; a chunk's probability of initial retrieval is determined by its degree of match with the retrieval cues. The retrieval probability of a chunk increases non-linearly with an increment in number of matching-cues implying that a chunk matching 1 out of 2 cues and a chunk

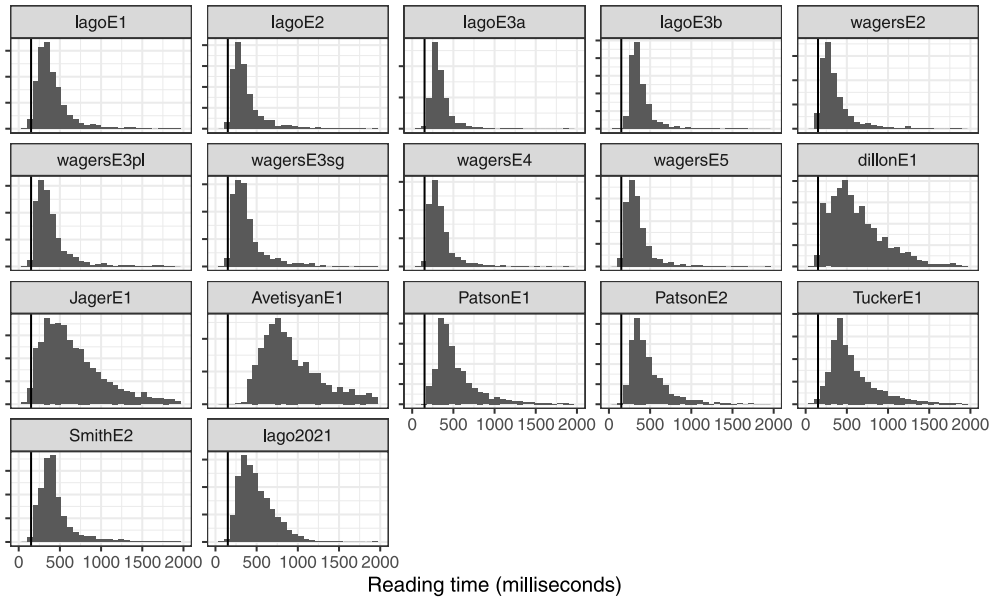


Fig. B.1. The distribution of reading times at the verb across subject–verb number agreement studies. The vertical line in each block is drawn at 150 ms.

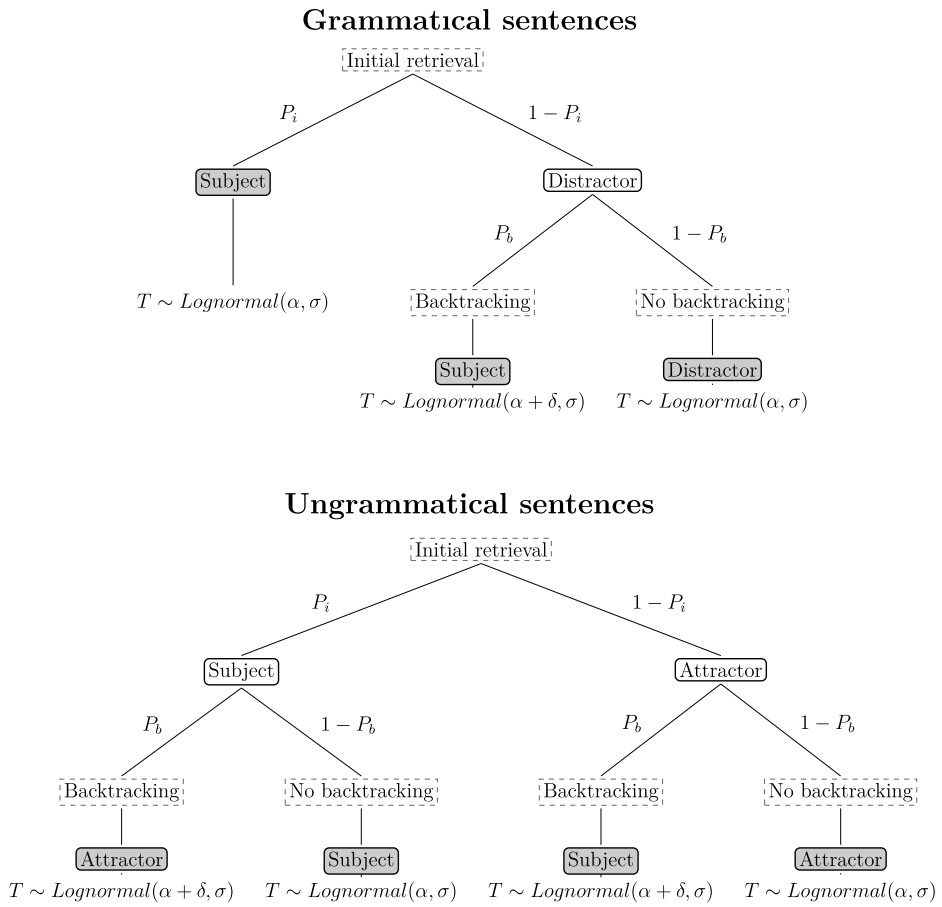


Fig. C.1. A schematic illustration of multinomial processing tree used to implement the non-linear cue-based retrieval model.

matching 0 out of 2 cues would have almost equal probability of initial retrieval. The retrieval probability of a memory chunk i is given by,

$$P_i = \frac{\prod_{j=1}^{N} S_{ji}^{W_j}}{\sum_{i=1}^{n} \prod_{j=1}^{N} S_{ji}^{W_j}} \tag{C.1}$$

where S_{ij} is the degree of match between a chunk i and a cue j, S_{ij} takes the value 0.99 if the chunk matches the cue and the value 0.01 when it does not; W_j is the weight of retrieval cue j. The initial retrieval takes a fixed time, say α, regardless of a chunk's degree of match with the retrieval cues.

The model further assumes that if the initially retrieved chunk does not fully match the retrieval cues, a backtracking can occur with some probability P_b. For example, in case of grammatical sentences, the backtracking occurs with probability P_b whenever the distractor is retrieved initially. Consequently, if the distractor was initially retrieved in some trials, these trials would increase the overall dependency completion time because of additional backtracking time δ.

However, the ungrammatical sentences pose an implementational challenge: Neither the subject nor the attractor fully match the retrieval cues, should backtracking occur for both the nouns or for only the attractor?

We present a model assuming that the backtracking can occur for the subject and the attractor noun in ungrammatical sentences.[19] To generate reading time predictions from the model, we use a multinomial processing tree following (Lissón, et al., 2021; Nicenboim & Vasishth, 2018). Fig. C.1 shows the decision tree of processing steps assumed in the model. The labels shown in the tree are as follows: α is the time taken (in log milliseconds) for initial retrieval, P_i is the probability of initial retrieval, δ is the time taken for backtracking (reanalysis of the incorrectly retrieved chunk).

Fig. C.1 shows that the reading times in both grammatical and ungrammatical sentences come from a mixture of two lognormal distributions - (i) when backtracking occurs, $Lognormal(\alpha + \delta, \sigma)$, (ii) when backtracking does not occur, $Lognormal(\alpha, \sigma)$.

In grammatical sentences, the probability of backtracking in the kth trial (regardless of what is retrieved initially) is $(1 - P_i) \cdot P_b$. Hence, the reading time in the kth trial can be sampled using

$$T_k \sim \begin{cases} Lognormal(\alpha + \delta, \sigma), & \text{if } z_k = 1 \\ Lognormal(\alpha, \sigma), & \text{if } z_k = 0 \end{cases}$$
$$\text{where } z_k \sim Bernoulli(\theta = (1 - P_i) \cdot P_b) \tag{C.2}$$

However, in the ungrammatical sentences, the backtracking can occur for both the subject and the attractor. The probability of backtracking in the kth trial (regardless of what is retrieved initially) is $P_i \cdot P_b + (1 - P_i) \cdot P_b = P_b$. Hence, the reading time in the kth trial

[19] The model assuming that backtracking occurs only for the non-subject (attractor/distractor) noun is also presented in this subsection; the model shows qualitatively opposite predictions with respect to observed effects in ungrammatical sentences.

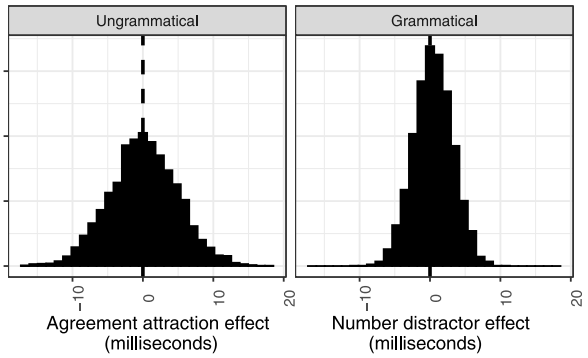


Fig. C.2. The agreement attraction and the number distractor effect predicted by the non-linear cue-based retrieval model.

can be sampled as

$$T_k \sim \begin{cases} Lognormal(\alpha+\delta,\sigma), & \text{if } z_k = 1 \\ Lognormal(\alpha,\sigma), & \text{if } z_k = 0 \end{cases} \quad \text{where } z_k \sim Bernoulli(\theta = P_b) \tag{C.3}$$

Using the above retrieval time equations, we derive predictions for the number agreement effects in grammatical and ungrammatical conditions for subject–verb agreement dependencies. Suppose, the effect in grammatical and ungrammatical conditions is E_g and E_u respectively. The effects come from the model conditional on its one free parameter, the scaling parameter α,

$$\begin{pmatrix} E_g \\ E_u \end{pmatrix} \sim Model(\alpha, \delta, P_i, P_b) \tag{C.4}$$

where P_i is the probability with which the subject noun is retrieved initially, Eq. (C.1) determines the value of P_i. The probability of backtracking P_b is sampled from a Beta prior $Beta(2,2)$. Finally, the scaling parameter α is assumed come from the following prior distribution (similar to the scaling parameter in the feature percolation model),

$$\alpha \sim Normal_{lb=5.2}(5.3, 0.05) \tag{C.5}$$

where $Normal_{lb=5.2}(5.3, 0.05)$ means a normal distribution with mean 5.3, standard deviation 0.05, and truncated at 5.2 so that it does not allow values lower than 5.2. The justification for the choice of priors on the scaling parameter is given in section Choice of priors on the scaling parameter .

Fig. C.2 shows the model predictions for the grammatical and ungrammatical subject–verb number agreement dependencies. In grammatical sentences, the probability of retrieving the subject noun is close to 100% in both conditions (a) and (b) (see Table C.1). As a result, the reading times in almost all the trials are sampled from $Lognormal(\alpha,\sigma)$ and hence, the model predicts close-to-zero effect in grammatical sentences along with some uncertainty. In ungrammatical sentences, the retrieval probability is 50%–50% for the subject and the attractor in condition (c) and approximately 100%–0% in condition (d) (see Table Table C.1). However, whatever is retrieved initially is subject to probabilistic backtracking. Consequently, in both conditions (c) and (d), reading times are sampled from a mixture of $Lognormal(\alpha,\sigma)$ and $Lognormal(\alpha+\delta,\sigma)$ depending on the probability of backtracking P_b. Thus, the model again predicts no difference between conditions (c) and (d) with some uncertainty (see Fig. C.2).

The non-linear cue-based retrieval model assuming that backtracking occurs only for the non-subject nouns

In the previous model, we have assumed that the backtracking can occurs for both the subject and the attractor in the ungrammatical sentences. We now turn to the second plausible scenario where backtracking occurs only for the distractor/attractor noun. In this model, the multinomial processing tree for the ungrammatical sentences will be same as for the grammatical sentences (see Fig. C.1). Consequently, the reading time equations will be the same for the grammatical and ungrammatical sentences:

$$T_k \sim \begin{cases} Lognormal(\alpha+\delta,\sigma), & \text{if } z_k = 1 \\ Lognormal(\alpha,\sigma), & \text{if } z_k = 0 \end{cases}$$
$$\text{where } z_k \sim Bernoulli(\theta = (1-P_i)\cdot P_b) \tag{C.6}$$

where $(1-P_i)\cdot P_b$ is the probability of backtracking, P_i is the probability of retrieving the subject noun initially, α is the time taken (in log milliseconds) in the initial retrieval, and δ is the time taken (in log ms) in the backtracking step.

Fig. C.3 shows the model predictions against the observed effects in grammatical and ungrammatical sentences. In grammatical sentences, since the probability of retrieving the subject noun is close to 100% in both conditions (a) and (b), the reading times are sampled from $Lognormal(\alpha,\sigma)$. Therefore, the model predicts close-to-zero effect in grammatical sentences. In ungrammatical sentences, the initial retrieval probability is 50%–50% for the subject and the attractor in condition (c) and approximately 100%–0% in condition (d) (see Table Table C.1). Consequently, in condition (c), the backtracking occurs in $0.5 \times P_b \times N$ trials out of total N trials, while in condition (d), the backtracking almost never occurs. This implies that the reading times in condition (c) are sampled from a mixture of $Lognormal(\alpha,\sigma)$ and $Lognormal(\alpha+\delta,\sigma)$, but in condition (d), they are sampled only from $Lognormal(\alpha,\sigma)$. Thus, the model predicts a slowdown in condition (c) compared to condition (d) in ungrammatical sentences; this prediction is inconsistent with the observed data (see Fig. C.3).

The non-linear cue-based retrieval model of Parker (2019)

Parker (2019) developed a non-linear cue-based retrieval model within the ACT-R cognitive architecture (Anderson, et al., 2004; Anderson & Lebiere, 2014). The model assumes that the dependency completion is driven by a cue-based search process in memory based on feature specifications such as [subject], [plural], called retrieval cues. Each chunk in memory receives activation via matching-cues in a non-linear (multiplicative) fashion. The total activation of a memory chunk i in trial k is given by,

$$A_{i,k} = B_i + \frac{\prod_{j=1}^{N} S_{ji}^{W_j}}{\sum_{i=1}^{n}\prod_{j=1}^{N} S_{ji}^{W_j}} + \epsilon_k \tag{D.1}$$

where B_i is the base-level activation of the chunk; S_{ij} is the strength of association between the chunk i and the retrieval cue j, S_{ij} takes the value 0.99 if the chunk matches the retrieval cue, and it takes the value 0.01 the chunk mismatches the retrieval cue. $\prod_{j=1}^{N} S_{ji}^{W_j}$ is the non-linear function that determines total activation received by the chunk from all the retrieval cues; ϵ_k is the trial-level noise in the activation such that $\epsilon_k \sim Normal(0,\sigma)$.

The model further assumes that a chunk with the highest total activation gets retrieved and the activation of the retrieved chunk determines the retrieval time at the verb. The retrieval time in the kth trial is given by

$$RT_k = Fe^{-A_{retr,k}} \tag{D.2}$$

where $A_{retr,k}$ is the activation of the retrieved chunk in the kth trial and F is a scaling parameter, called the latency factor. The latency factor F is the only free parameter in the model which is assumed to come from the following prior distribution

$$F \sim Normal_{lb=0.05}(0.15, 0.05) \tag{D.3}$$

Based on the above prior, we derive predictions for the number agreement effects in grammatical and ungrammatical sentences.

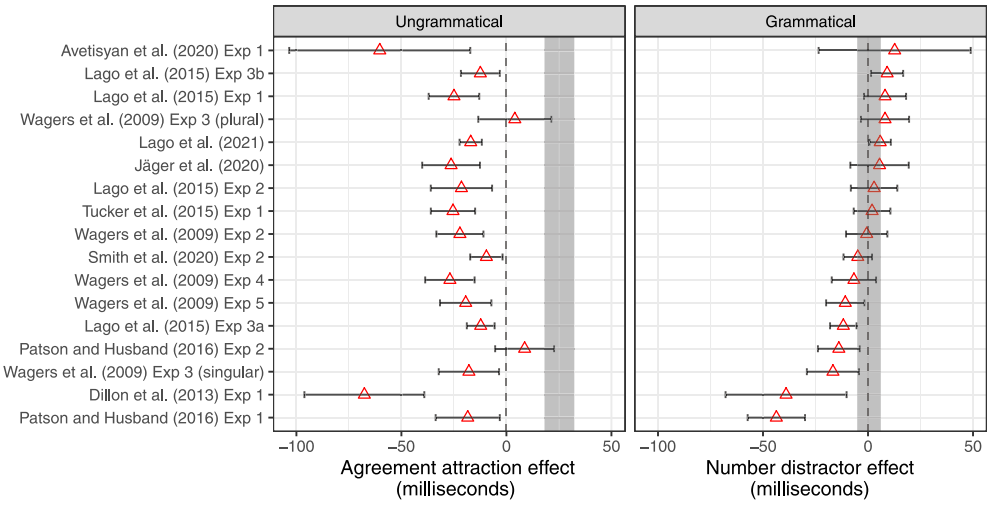


Fig. C.3. Prior predictions of the non-linear cue-based retrieval model assuming that backtracking occurs only for the attractor/distractor noun: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model. The red triangles and the error bars around them show the observed effect for each dataset specified on the *y*-axis.

Table C.1
The probability of initial retrieval in four conditions; P_i indicate the probability of retrieving the subject noun; 0.99 is the value of match between a chunk and a retrieval cue, 0.01 is the value of mismatch between a chunk and a retrieval cue.

Condition	The probability of retrieving the Subject noun P_i	Distractor/Attractor noun $(1 - P_i)$
(a) Grammatical, singular distractor condition		
The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{+singular}$ $was^{subject}_{singular}$ rusty.	$\frac{0.99\times0.99}{0.99\times0.99+0.99\times0.01} = 0.990$	$1 - 0.990 = 0.010$
(b) Grammatical, plural distractor condition		
The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{-singular}$ $was^{subject}_{singular}$ rusty.	$\frac{0.99\times0.99}{0.99\times0.99+0.01\times0.01} = 0.999$	$1 - 0.999 = 0.001$
(c) Ungrammatical, plural attractor condition		
The $key^{+subject}_{-plural}$ to the $cabinets^{-subject}_{+plural}$ $were^{subject}_{plural}$ rusty.	$\frac{0.99\times0.01}{0.99\times0.01+0.99\times0.01} = 0.500$	$1 - 0.500 = 0.500$
(d) Ungrammatical, singular attractor condition		
The $key^{+subject}_{-plural}$ to the $cabinet^{-subject}_{-plural}$ $were^{subject}_{plural}$ rusty.	$\frac{0.99\times0.01}{0.99\times0.01+0.01\times0.01} = 0.990$	$1 - 0.990 = 0.010$

Table D.1
The activation received by the subject and the distractor/attractor noun; 0.99 indicate the value of match between a chunk and a retrieval cue, 0.01 indicate the value of mismatch between a chunk and a retrieval cue.

Condition	Activation received by the Subject noun	Distractor/Attractor noun
(a) Grammatical, singular distractor condition		
The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{+singular}$ $was^{subject}_{singular}$ rusty.	$\frac{0.99\times0.99}{0.99\times0.99+0.99\times0.01} = 0.990$	$1 - 0.990 = 0.010$
(b) Grammatical, plural distractor condition		
The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{-singular}$ $was^{subject}_{singular}$ rusty.	$\frac{0.99\times0.99}{0.99\times0.99+0.01\times0.01} = 0.999$	$1 - 0.999 = 0.001$
(c) Ungrammatical, plural attractor condition		
The $key^{+subject}_{-plural}$ to the $cabinets^{-subject}_{+plural}$ $were^{subject}_{plural}$ rusty.	$\frac{0.99\times0.01}{0.99\times0.01+0.99\times0.01} = 0.500$	$1 - 0.500 = 0.500$
(d) Ungrammatical, singular attractor condition		
The $key^{+subject}_{-plural}$ to the $cabinet^{-subject}_{-plural}$ $were^{subject}_{plural}$ rusty.	$\frac{0.99\times0.01}{0.99\times0.01+0.01\times0.01} = 0.990$	$1 - 0.990 = 0.010$

Fig. D.1 shows the prior predictions of the model against the observed data. In grammatical sentences, the model predicts close-to-zero effect along with some uncertainty, a typical prediction of the non-linear cue combination. In case of ungrammatical sentences, both the subject and the attractor nouns receive same amount of activation (i.e. 0.5) in condition (c), but in condition (d) the subject receives most of the activation that is 0.99 (see Table D.1). Consequently, the chunk retrieved in condition (c) would have an approximate activation of 0.5 while the retrieved chunk in (d) would mostly have activation of approx. 0.99. Therefore, the model predicts a slowdown in condition (c) compared to condition (d). The model is thus able to capture the observed effects in grammatical sentences, but fails to capture the agreement attraction effects in ungrammatical sentences (see Fig. D.1).

The non-linear cue-based retrieval model of Parker (2019) assuming activation sampling

The previous model based on Parker (2019) assumes that activation received by a chunk is normalized by the sum of activations received by all the chunks in memory i.e. $\frac{\prod_{j=1}^{N} s_{ji}^{W_j}}{\sum_{i=1}^{n}\prod_{j=1}^{N} s_{ji}^{W_j}}$, and the chunk with highest activation is retrieved. However, we can make slightly different assumptions about how cue matching affect retrieval:

1. Each chunk in memory has a certain probability of retrieval based on their degree of match with the retrieval cues. The retrieval probability of a chunk i is given by

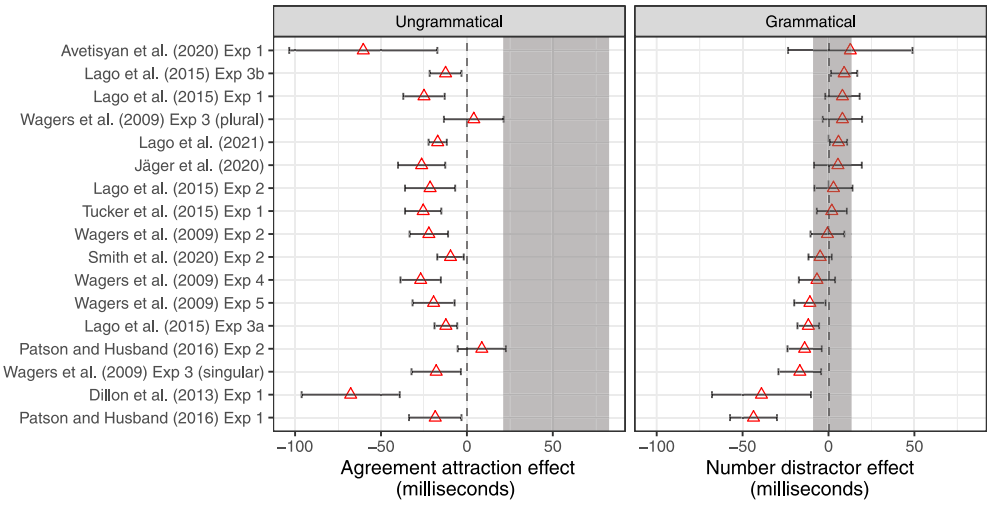


Fig. D.1. Prior predictions of the non-linear cue-based retrieval model of Parker (2019): The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model. The red triangles and the error bars around them show the observed effect for each dataset specified on the y-axis.

Table D.2
The activation and the retrieval probabilities of the subject and the distractor/attractor noun.

Condition	Activation received by the Subject	Activation received by the Distractor/Attractor	Retrieval probability of the Subject	Retrieval probability of the Distractor/Attractor
(a) Grammatical, singular distractor condition				
The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{+singular}$ $was^{subject}_{singular}$ rusty.	0.9801	0.0099	0.990	0.010
(b) Grammatical, plural distractor condition				
The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{-singular}$ $was^{subject}_{singular}$ rusty.	0.9801	0.0001	0.999	0.001
(c) Ungrammatical, plural attractor condition				
The $key^{+subject}_{-plural}$ to the $cabinets^{-subject}_{+plural}$ $were^{subject}_{plural}$ rusty.	0.0099	0.0099	0.500	0.500
(d) Ungrammatical, singular attractor condition				
The $key^{+subject}_{-plural}$ to the $cabinet^{-subject}_{-plural}$ $were^{subject}_{plural}$ rusty.	0.0099	0.0001	0.990	0.010

$$\theta_i = \frac{\prod_{j=1}^{N} s_{ji}^{w_j}}{\sum_{i=1}^{n} \prod_{j=1}^{N} s_{ji}^{w_j}}$$

The retrieval probabilities associated with the chunks in memory determine which of them is retrieved in a trial.

2. The activation of a chunk i is given by
$A_i = \prod_{j=1}^{N} s_{ji}^{w_j}$
3. In the kth trial, one of the chunks from memory is retrieved conditional on their respective probabilities of retrieval. For example, if there are two chunks in memory with activations A_1 and A_2 and retrieval probabilities θ_1 and $1-\theta_1$, the activation of the retrieved chunk in the kth trial can be derived as $A_{retr,k} = \begin{cases} A_1, & \text{if } z_k = 1 \\ A_2, & \text{if } z_k = 0 \end{cases}$ where $z_k \sim Bernoulli(\theta_1)$
4. Finally, the retrieval time in the kth trial is determined by the activation of the retrieved chunk
$T_k = Fe^{-f A_{retr,k}}$

Fig. D.2 shows the prior predictions of the Parker (2019) model assuming retrieval probability-based sampling of nouns for dependency completion. In grammatical sentences, the subject nouns has almost 100% probability of retrieval in both conditions (a) and (b). Therefore, the retrieved chunk in both (a) and (b) would almost always have the same activation of 0.9801 (see Table D.2). Therefore, the model predicts no difference between conditions (a) and (b) consistent with the observed pattern of effects. In ungrammatical sentences, the retrieval probability is 50%–50% in condition (c), therefore, the retrieved chunk would always have the activation 0.0099. In condition (d), the subject is retrieved most of the time, but it also has the same activation 0.0099 (see Table D.2). Therefore, the model again predicts no difference between conditions (c) and (d). This prediction is inconsistent with the observed agreement attraction effects in ungrammatical sentences (see Fig. D.2).

The feature percolation model

The feature percolation model assumes that the number feature of the noun which is not the subject of the verb percolates to the subject noun in θ proportion of trials. For example, consider the sentence *The key to the cabinets were rusty*. The plural feature of the noun phrase *the cabinets* percolates up to the subject, *the key* in $\theta \times N$ number of trials out of total N trials; this changes the representation of subject from singular to plural in $\theta \times N$ trials. Consequently, the subject will match in number feature with the verb in $\theta \times N$ trials, and will not match in the remaining $(1-\theta) \times N$ trials. The model further assumes that the processing at the verb is faster when the subject and the verb match in number feature compared to when they do not. The reading times at the verb thus come from different distributions in these two situations: when the subject matches the verb in the number feature and when it does not.

Suppose that when the subject matches the verb in the number feature, the reading times x at the verb come from a probability density function $p_1(x|\zeta)$; and, when the subject mismatches the verb in number, the reading times x come from a probability density function $p_2(x|\zeta)$; where ζ is the vector of parameters in the model. As we know for the above sentence that the subject matches the verb in θ proportion of trials, we can say that the reading times are generated from $p_1(x|\zeta)$ with a probability of θ and from $p_2(x|\zeta)$ with a probability of $(1-\theta)$. We can express this process using a two-component (finite) mixture

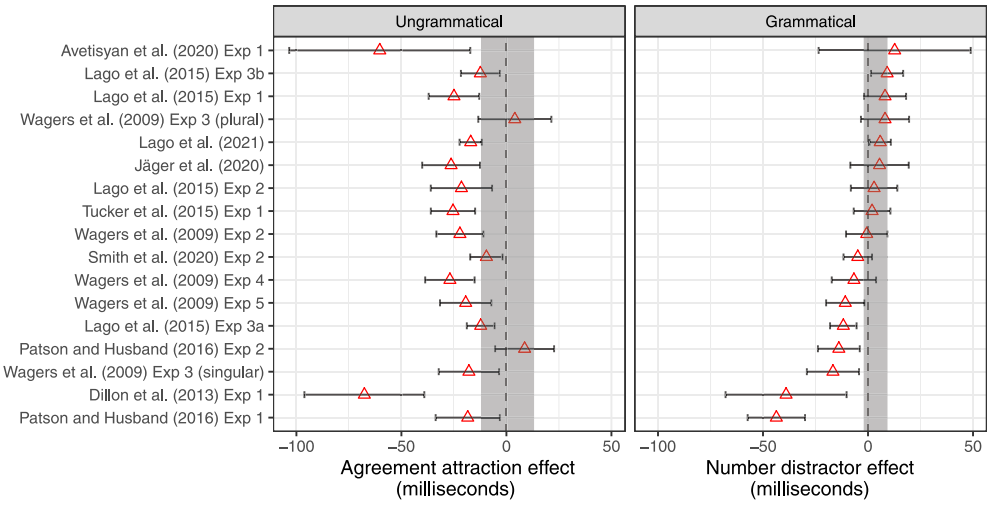


Fig. D.2. Prior predictions of the non-linear cue-based retrieval model of Parker (2019) assuming activation sampling: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model. The red triangles and the error bars around them show the observed effect for each dataset specified on the y-axis.

model (Frühwirth-Schnatter, 2006; McLachlan & Peel, 2004). The reading times x come from a mixture of probability density functions $p_1(x|\zeta)$ and $p_2(x|\zeta)$ with probabilities θ and $(1-\theta)$ respectively. The distribution of reading times over all the trials can be represented by a density function $f(x|\zeta)$ such that

$$f(x|\zeta) = \begin{cases} p_1(x|\zeta), & \text{with probability } \theta \\ p_2(x|\zeta), & \text{with probability } (1-\theta) \end{cases} \tag{E.1}$$

The reading time at the verb in the ith trial, RT_i will be

$$RT_i \sim \begin{cases} p_1(x|\zeta), & \text{if } z_i = 1 \\ p_2(x|\zeta), & \text{if } z_i = 0 \end{cases} \text{ where } z_i \sim Bernoulli(\theta) \tag{E.2}$$

For $p_1(x|\zeta)$ and $p_2(x|\zeta)$, we choose lognormal distributions with means μ_{match} and $\mu_{mismatch}$ respectively, and the same standard deviation σ. The rationale behind choosing the lognormal distributions is that the reading times from the self-paced reading experiments can be modeled as log-normally distributed: the reading times can only take positive real values and tend to have a long tail of relatively large values (Nicenboim et al., 2022). The reading time is thus assumed to come from a mixture of two lognormal distributions.

$$RT_i \sim \begin{cases} Lognormal(\mu_{match}, \sigma), & \text{if } z_i = 1 \\ Lognormal(\mu_{mismatch}, \sigma), & \text{if } z_i = 0 \end{cases} \text{ where } z_i \sim Bernoulli(\theta) \tag{E.3}$$

As the model assumes that the reading times in the subject–verb number match situation are faster than in mismatch situation, we can write $\mu_{mismatch}$ as

$$\mu_{mismatch} = \mu_{match} + d \tag{E.4}$$

where d is constrained be positive, $d > 0$. Let us call μ_{match} the scaling parameter and represent it as S. We can rewrite the reading time function as,

$$RT_i \sim \begin{cases} Lognormal(S, \sigma), & \text{if } z_i = 1 \\ Lognormal(S + d, \sigma), & \text{if } z_i = 0 \end{cases} \text{ where } z_i \sim Bernoulli(\theta) \tag{E.5}$$

where the scaling parameter, S and the percolation rate θ are assumed to be the free parameters, with the following priors

$$S \sim Normal_{lb=5.2}(5.3, 0.05),$$

and

$$\theta \sim Normal_{lb=0.1}(0, 0.25),$$

Here, $lb = 0.1$ indicates a lower bound on distortion rate values. The lower bound of 0.1 implies that the distortion rate lies in the range [0.1, 0.5] — that is, the representation of the subject noun changes in at least 10% and at most 50% of the trials.

The marking and morphing model

The Marking and Morphing model (Eberhard et al., 2005) generates almost the same predictions as the feature percolation model; although, unlike the feature percolation model, the marking and morphing model predicts an asymmetry in the effect sizes for grammatical vs. ungrammatical sentences. The model assumes that

1. The nouns in memory have continuous-valued number information such that an unequivocally singular noun would have the lowest value and an unequivocally plural noun would have the highest value.
2. The number information for the subject noun comes from two sources, the number marking on the subject and the lexical number specifications of the subject and the local nouns.
3. The reading time at the verb is determined by the degree of number match between the subject and the verb.

Consider the sentence *the key to the cabinets was rusty*. The number value of the subject noun *the key*, $S(r)$ is determined by (a) the number marking of the subject, $S(n)$, (b) lexical number specification of the subject noun, $S(m_1)$, and (c) lexical number specification of the distractor noun *the cabinets*, $S(m_2)$,

$$S(r) = S(n) + \sum_j W_j \cdot S(m_j) \tag{F.1}$$

where $S(m_j)$ represent the lexical number specification of the jth noun, W_j represent the weights for feature transmission from jth noun. Here, $S(m_j)$ has value 0 if the jth noun is a singular count noun, similarly it has value 1.16 for the plural count noun, 1 for the invariant plural noun, 0.09 for the singular collective noun, and 1.25 for the plural collective noun.

Suppose the weight for the subject noun is W_1 and weight for the local noun is W_2, the equation can be rewritten as

$$S(r) = S(n) + W_1 \cdot S(m_1) + W_2 \cdot S(m_2) \tag{F.2}$$

The above valuation of plurality of the subject noun can be used to determine the probability of producing/predicting a plural verb using $P(plural) = \frac{1}{1+e^{-(S(r)+b)}}$ where b is a constant fixed to a value of -3.42. But we are interested in predicting reading times at the verb.

We need a linking function to derive reading time predictions from the model. In order to maintain the continuous number valuation property of the marking and morphing model, we compute the degree of match between the subject noun and the verb on a continuous scale. The degree of match in plurality of the verb and the subject noun is given by

$$PL_{match} = \left| \frac{S(r)}{mp} - V_{pl,k} \right| \tag{F.3}$$

where mp – maximum plurality – is a normalizing constant to scale the value of $S(r)$ on a scale of 0–1; mp takes the value of maximum possible plurality the subject noun can have; $V_{pl,k}$ indicate the plurality of the verb in trail k such that the value of V_{pl} is 1 if the verb is plural and 0 if the verb is singular.

The degree of match PL_{match} is multiplied with a constant δ to compute the number mismatch cost Δ_k on log milliseconds scale:

$$\Delta_k = \left| \frac{S(r)}{mp} - V_{pl,k} \right| \cdot \delta \tag{F.4}$$

The mismatch cost Δ_k can have values on a continuous scale between 0 and δ; 0 if the subject number completely match the verb number and δ if the subject number completely differs from the verb number.

The reading times at the verb in the kth trial come from a lognormal distribution,

$$RT_k \sim Lognormal(\alpha + \Delta_k, \sigma) \tag{F.5}$$

where α is the scaling parameter, it can be interpreted as the mean reading time (in log ms) when the subject noun completely match the verb in number, Δ_k is the number-mismatch penalty in the kth trial computed using Eq. (G.1). We assume the scaling parameter α and the weight for local noun W_2 are two free parameters in the model. The scaling parameter has the following prior,

$$\alpha \sim Normal_{lb=5.2}(5.3, 0.05),$$

For W_2, we set a prior on θ such that $\theta = W_2/k$ (where k is fixed at 9), θ will determine the rate of number feature spread from the local noun to the subject noun:

$$\theta \sim Normal_{lb=0.1}(0, 0.25),$$

where $lb = 0.1$ indicate a lower bound of 0.1 on the distortion rate.

Given the above priors, the prior predictions of the model are shown in Fig. G.1.

The grammaticality bias model

Following Hammerly et al. (2019)'s idea of response bias, we assume a similar bias in reading such that the comprehender has a strong expectation to encounter grammatical continuation of a partially-read sentence. For example, after the participants have read *the key to the cabinets...*, they have a bias to expect a singular verb phrase. When a singular verb is encountered, any change in continuous-valued number of the subject does not cause much difference in processing at the verb. But when a plural verb is encountered, the processing is already difficult, and as a result, any slight evidence of plurality in subject's number causes a facilitation at the verb. To operationalize this idea, we use an exponential *mismatch cost function*: As the degree of number-mismatch between the verb and the subject increases, the processing cost at verb increases exponentially. We can implement the grammaticality bias model within the framework of the marking and morphing mechanism (section 'The Marking and Morphing model').

The grammaticality bias model replaces the linear mismatch cost (see Eq. (G.1)) of the marking and morphing model by an exponential mismatch cost function. The mismatch cost in trial k for the grammaticality bias model is given by

$$\Delta_k = \left(\left| \frac{S(r)}{mp} - V_{pl,k} \right| \cdot \delta \right)^{2b} \tag{G.1}$$

where b is the grammaticality bias and it can take values between 0.5 and 1 such that $b = 0.5$ corresponds to no bias (default marking and morphing model) and $b = 1$ means the maximum bias; mp – maximum plurality – is a normalizing constant to scale the value of $S(r)$ on a scale of 0–1; $V_{pl,k}$ indicate the plurality of the verb in trail k such that the value of V_{pl} is 1 if the verb is plural and 0 if the verb is singular.

The mismatch penalty Δ_k can now have values between 0 and δ^2; 0 if the subject number completely match the verb number and δ^2 if the subject number completely differs from the verb number and the grammaticality bias b is equal to 1. The mismatch cost exponentially increases with increase in degree of mismatch when the grammaticality bias is greater than 0.5. For higher grammaticality bias, the mismatch cost has more exponential growth.

The reading time at the verb in the kth trial is assumed to come from a lognormal distribution

$$RT_k \sim Lognormal(\alpha + \Delta_k, \sigma) \tag{G.2}$$

where α is the scaling parameter. Using the same priors as in the marking and morphing model, we derive model predictions for the grammatical and ungrammatical sentences. Fig. F.1 shows the prior predictions of the model. Why does the grammaticality bias reduce the effect size in grammatical sentences? This is illustrated is Fig. G.2. The mismatch cost function is an exponential curve; the grammatical sentences that are on the lower side of the curve would have very small influence of the plurality of the subject compared to ungrammatical sentences which are on the higher side of the curve.

The lossy compression model

The lossy compression model assumes that the comprehender has access to only an imperfect memory representation of the linguistic input and the processing difficulty at a verb is the expected surprisal of seeing the verb over all possible memory representations of the preverbal input. For example, consider the sentence *The key to the cabinets was rusty*, The input here is $I = N \quad P \quad N.pl$ where N represents a noun, P represents a preposition, and $.pl$ represents a plural marker on a noun.

The input I gets distorted to a possible memory representations r_i such that the plural maker on a noun is inserted or deleted. The following memory representations are possible,

$r_1 = N.pl \quad P \quad N.pl \quad r_2 = N.pl \quad P \quad N$

$r_3 = N \quad P \quad N.pl \quad r_4 = N \quad P \quad N$

The processing difficulty for the upcoming verb V will be the *expected surprisal* of the verb given all possible memory representations $r_1, r_2, \ldots, r_N$:

$$D(V|I) = \sum_{i=1}^{N} -\log P(V|r_i) \cdot P(r_i|I) \tag{H.1}$$

where $P(r_i|I)$ is the probability of obtaining a memory representation r_i from the actual pre-verbal input I. And, $-\log P(V|r_i)$ is the surprisal – negative log conditional probability – of seeing a plural/singular verb given a memory representation r_i. The model further assumes that the comprehender reconstructs a set of possible, true preverbal contexts from their memory representation r_i based on their prior linguistic knowledge and their uncertainty about the degree of distortion in the system. We can derive the conditional probability $P(V|r_i)$ by marginalizing out the all possible true contexts $c_1, c_2, \ldots, c_n$

$$P(V|r_i) = \sum_{j=1}^{n} P(V|c_j)P(c_j|r_i) \tag{H.2}$$

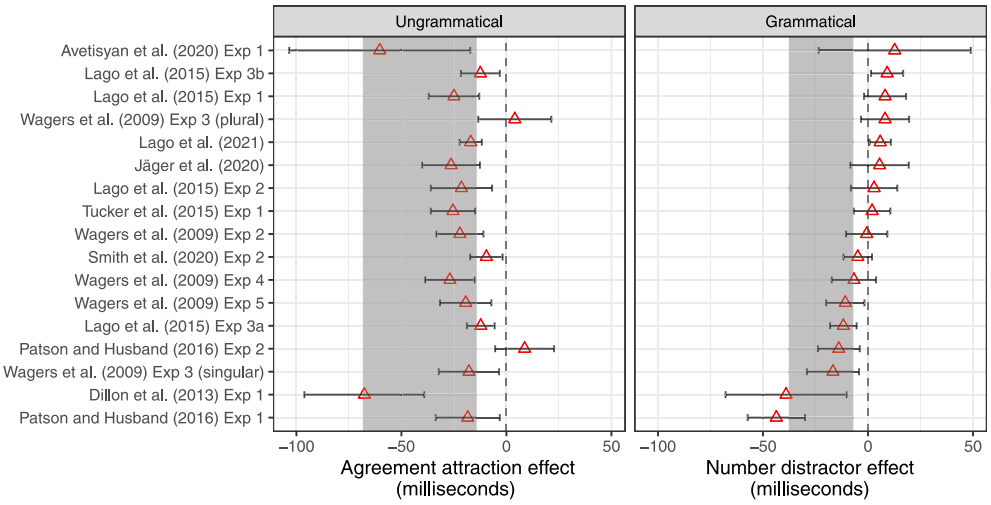


Fig. F.1. Prior predictions of the Marking and Morphing model: The shaded gray bands represent the 95% credible intervals of number agreement effects predicted by the model. The red triangles and the error bars around them show the observed effect for each dataset specified on the y-axis.

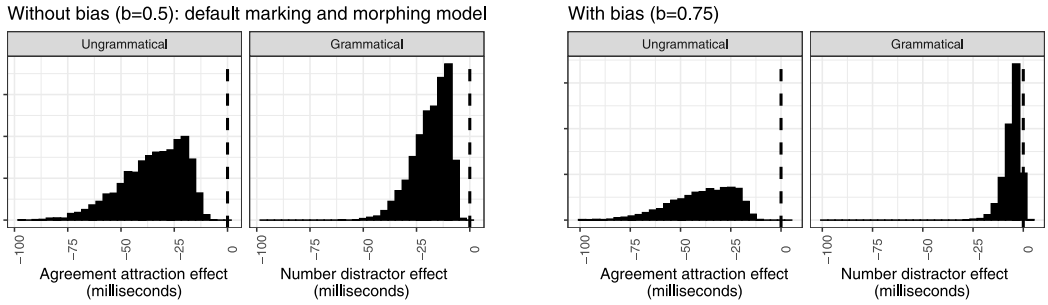


Fig. G.1. The agreement attraction and the number distractor effect predicted by the grammaticality bias model.

where $P(V|c_j)$ is the conditional probability of seeing the verb given a possible true context c_j. We computed conditional probabilities from the COW corpora (Schäfer, 2015; Schäfer & Bildhauer, 2012) (for English and Spanish studies) and the Universal Dependencies treebanks (Nivre et al., 2018) (for Arabic and Armenian). The search criteria used for computing probabilities from the corpora were as follows. Fully structural queries were used to look for each construction. For example, to search the prepositional phrase (PP) constructions, a noun phrase followed by an embedded PP was set as a probe. Similarly, for searching relative clause (RC) constructions, the probe was: a noun phrase followed by an RC that starts with a relative pronoun and contains a subject noun modifying a finite verb. The information used in the queries include Part-of-Speech tags, verb type (finite vs. non-finite), dependency relation, and number (singular vs. plural). We do not use lexical information from the corpora, implying that we do not make distinction between different types of prepositions and relative pronouns.

We can derive the probability $P(c_j|r_i)$ up to proportionality using Bayes' rule,

$$P(c_j|r_i) \propto \mathcal{L}(r_i|c_j)P(c_j) \tag{H.3}$$

where $P(c_j)$ represents the probability of seeing the representation c_j in the corpus, and $\mathcal{L}(r_i|c_j)$ is the likelihood of generating the memory representation r_i from a possible true representation c_j.

Based on Eqs. (H.2) and (H.3), we can rewrite the processing difficulty function as follows,

$$D(V|I) = \sum_{i=1}^{N} -\log P(V|r_i) \cdot P(r_i|I) \tag{H.4}$$

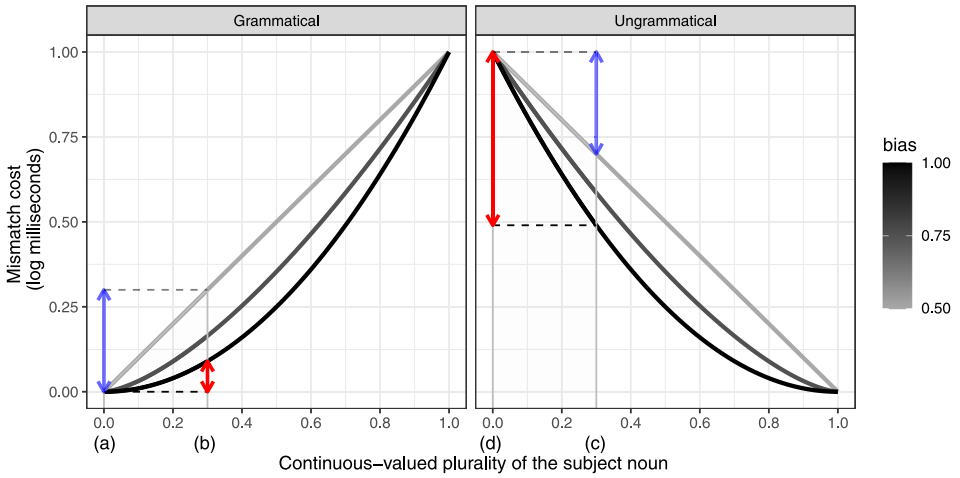
where

$$P(V|r_i) \propto \sum_{j=1}^{n} P(V|c_j)\mathcal{L}(r_i|c_j)P(c_j)$$

The likelihood function $\mathcal{L}(r_i|c_j)$ is called the lossy memory encoding function: the likelihood that a true representation c_j gets distorted to memory representation r_i given a deletion rate d and insertion rate a (see Table H.1).[20]

$$r_i|c_j \sim Memory(d, a) \tag{H.5}$$

[20] The same memory function also underlies $P(r_i|I)$: the probability of generating a memory representation r_i from the observed linguistic input I. The probability $P(r_i|I)$ can be calculated from insertion and deletion rates in the same way as we calculate $\mathcal{L}(r_i|c_j)$ (see Table H.1). The function $P(r_i|I)$ represents the experimenter's uncertainty about the memory representation formed by the comprehender and the function $\mathcal{L}(r_i|c_j)$ represents the comprehender's uncertainty about the true intended representation.



(a) The $key^{+subject}_{+singular}$ to the $cabinet^{-subject}_{+singular}$ $was^{subject}_{singular}$ rusty

(b) The $key^{+subject}_{+singular}$ to the $cabinets^{-subject}_{-singular}$ $was^{subject}_{singular}$ rusty

(c) * The $key^{+subject}_{-plural}$ to the $cabinets^{-subject}_{+plural}$ $were^{subject}_{plural}$ rusty

(d) * The $key^{+subject}_{-plural}$ to the $cabinet^{-subject}_{-plural}$ $were^{subject}_{plural}$ rusty

Fig. G.2. An illustration of number-mismatch cost, Δ at the verb as a function of subject's plurality and the grammaticality bias. When there is no grammaticality bias (i.e., bias = 0), the mismatch cost increases linearly w.r.t. plurality of the subject but as the value of bias increases, the mismatch cost grows exponentially. The labels (a), (b), (c), and (d) on the x-axis mark the number value of the subject noun in four conditions shown below the graph. The red arrows – indicating difference in mismatch costs between a pair of conditions – can be interpreted as the attraction effects on the log scale when there is a grammaticality bias ($b = 1$); similarly, the blue arrows indicate attraction effects (on log scale) when there is no bias ($b = 0.5$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table H.1
The lossy memory encoding function: the likelihood of obtaining the memory representation r_i from the distortion of a possible true representation (N P $N.pl$).

Memory representation	Likelihood of obtaining r_i from $c_j = N$ P $N.pl$
r_i	$\mathcal{L}(r_i \mid c_j)$
$N.pl$ P $N.pl$	$a(1-d)$
$N.pl$ P N	ad
N P $N.pl$	$(1-a)(1-d)$
N P N	$(1-a)d$

where d is the rate of deleting a plural marker and a is the rate of inserting a plural marker. Table H.1 shows the likelihood of obtaining memory representation r_i from a possible true representation $c_j = N$ P $N.pl$.

Finally, we transform processing difficulty into reading times using a linear linking function. Reading times in kth trial, RT_k, will be:

$$RT_k = A + S \cdot D(V|I) + \epsilon_k \tag{H.6}$$

where S is a scaling parameter and ϵ_k is the random noise in the kth trial such that $\epsilon_k \sim Normal(0, 20)$; the parameter A is the intercept of the linear function and represents the shift in reading times, we keep A fixed at 120 ms.

Thus, the model has three free parameters: scaling parameter S, deletion rate d, and insertion rate a, with the following priors,

$$S \sim Normal_{lb=0.15}(0.25, 0.05)$$

$$d \sim Normal_{lb=0.1}(0, 0.25)$$

$$a \sim Normal_{lb=0.1}(0, 0.25)$$

where $lb = 0.1$ is the lower bound on the deletion rate and insertion rate values. The parameters a and d represent the rate of information loss when the linguistic input is stored in memory.

The feature percolation-plus-retrieval model

The feature percolation-plus-retrieval model assumes that the number feature of the non-subject noun probabilistically percolates to the subject noun which changes the subject's representation in a proportion of trials before the retrieval is triggered at the verb. For example, consider the sentence *The key to the cabinets was rusty*. The plural feature of *the cabinets* percolates up to the subject noun phrase *the key* with a probability θ. Suppose there are total N trials in an experiment. In $\theta \times N$ trials, the subject noun would now have plural feature and in remaining $(1-\theta) \times N$ trials it would have the singular feature (see Fig. I.1). Thus, the representation of nouns in ith trial, say r_i, is a function of percolation probability, θ and the original preverbal input I,

$$r_i \sim f(I, \theta) \tag{I.1}$$

The probability of feature percolation θ is called a *distortion rate parameter* as it determines the rate of change in representation of the preverbal input in our implementation.

The model further assumes that each noun phrase would receive activation based on their degree of match with the retrieval cues. For example, if the retrieval cues at verb are [subject], [singular], then a singular subject noun would receive more activation compared to a plural subject noun. Therefore, the amount of activation received by each noun phrase depends on their feature representation (see Fig. I.1 for the schematic illustration). The activation of noun phrases n_1 and n_2 in the ith trial is given by

$$\begin{pmatrix} A_{n_1,i} \\ A_{n_2,i} \end{pmatrix} = Activation(r_i) + \epsilon_i \tag{I.2}$$

where $Activation$ represent the activation function that determines the activation received by a chunk based on its degree of match with the retrieval cues, r_i is the representation of the nouns in the ith trial (see Eq. (I.1)), and ϵ_i is the trial-level Gaussian noise in the activation.

A chunk with the highest activation gets retrieved at the verb and the retrieval times are determined by the activation of the retrieved

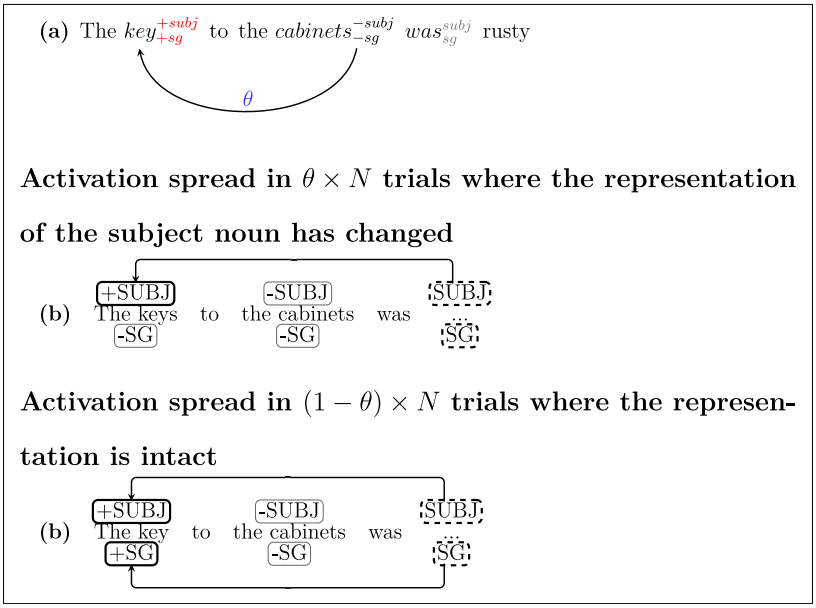


Fig. I.1. Schematic illustration of activation received by the nouns in the feature percolation-plus-retrieval model: Feature percolation modulates the amount of activation received by the subject noun during the retrieval process. In $\theta \times N$ trials, the subject noun receives less activation in condition (b) compared to condition (a).

chunk. The retrieval times at the verb in the ith trial is given by

$$RT_i = Fe^{-max(A_{n_1,i}, A_{n_2,i})} \tag{I.3}$$

where $max(A_{n_1,i}, A_{n_2,i})$ represent the maximum of the activation of the nouns in the ith trial. F is the scaling parameter.

The model has two free parameters, the distortion rate θ and the scaling parameter F, with the following priors. On the distortion rate θ, we choose the same truncated normal priors as we did for distortion rate parameters in other representation distortion-based models.

$$\theta \sim Normal_{lb=0.1}(0, 0.25) \tag{I.4}$$

where $lb = 0.1$ represents the lower bound of 0.1 on distortion rate values. For the scaling parameter F, which comes from retrieval-part of the model, we choose the same prior as for latency factor in cue-based retrieval model,

$$F \sim Normal_{lb=0.05}(0.15, 0.05) \tag{I.5}$$

The lossy compression-plus-retrieval model

The lossy compression-plus-retrieval model assumes that when the preverbal linguistic input is stored in memory, it gets distorted probabilistically to an imperfect memory representation before the retrieval is triggered at the verb. For example, in the sentence *the key to the cabinets was rusty*, the preverbal input is $NPN.pl$ meaning that the first noun is singular and the second noun is plural. When stored in memory, this input can get distorted to a memory representation $N.plPN$ such the plural marker is inserted at the first noun and deleted from the second noun. Thus, the likelihood of obtaining a memory representation r from an input I is a function of the rate of deleting plural markers d, and the rate of inserting plural markers on nouns a. In the jth trial, the input I transforms to memory representation r_j, such that,

$$r_j|I \sim f(a, d) \tag{J.1}$$

For parameters a and d, we choose the same priors as for insertion and deletion rates in the lossy compression model (see section 'The lossy compression model').

The lossy-compression-plus-retrieval model assumes that the chunk with the highest activation gets retrieved in each trial. And, the activation of a chunk is determined by the amount of activation it receives based on the degree of match between the chunk's features and retrieval cues. For example, a singular subject noun would receive more activation from [subject], [singular] cues at the verb compared to a plural subject noun. The activation of the retrieved chunk is, therefore, a function of memory representation of preverbal noun phrases in the jth trial,

$$A_{j,retrieved} \sim Activation(r_j) \tag{J.2}$$

As discussed in section 'The cue-based retrieval model', the retrieval time at the verb in the jth trial, RT_j, is an exponential function of the activation of the retrieved chunk in that trial,

$$RT_j = Fe^{-A_{j,retrieved}} \tag{J.3}$$

where F is a scaling parameter called the latency factor which reflects overall processing time; $A_{j,retrieved}$ is the activation of chunk retrieved in trial j. We choose same prior on the scaling parameter F as on the scaling parameter in the cue-based retrieval model (see section 'The cue-based retrieval model').

Parameter estimation

A key step in model evaluation is parameter estimation. One needs to estimate what values of the parameter, say θ, of the model would have generated the observed data y. Under the Bayesian approach, we can estimate the posterior distribution of parameter values given the observed data, $\pi(\theta|y)$, using Bayes' rule

$$\pi(\theta|y) = \frac{\mathcal{L}(\theta|y)\pi(\theta)}{\pi(y)} \tag{K.1}$$

where $\mathcal{L}(\theta|y)$ is the *likelihood function*, i.e., the probability density of obtaining the data for given parameter value; $\pi(\theta)$ is the prior distribution of θ, which represents the prior knowledge about θ before the data is available; $\pi(y)$ is the marginal likelihood of obtaining the data y taken over all possible parameter values.

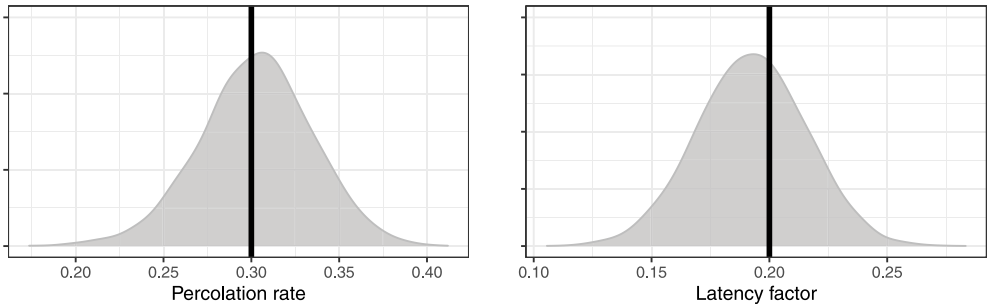


Fig. K.1. Estimated posterior distributions for the latency factor and the percolation rate are shown using gray density graphs and true parameter values are shown as black vertical lines.

Posterior simulation methods, like Markov Chain Monte Carlo, use the likelihood function and the prior density function to estimate the posterior distribution of the parameter. But for the models that we evaluated here, parameter estimation is challenging because the likelihood function cannot be expressed analytically for most of the models, at least not without considerable simplification. This is one reason why we do not use the standard Bayesian estimation methods but rather use Approximate Bayesian Computation. This method has the advantage that we can extend the approach that we used in the present paper to even more complex models, such as the extended SWIFT model of eye-movement control and reading (Engbert, et al., 2022; Rabe, et al., 2021) and recent implementations of the SOPARSE model (Smith et al., 2021; Smith & Vasishth, 2022).

Approximate Bayesian Computation (ABC) is a well-established and effective tool for parameter estimation (Palestro, et al., 2018; Sisson et al., 2018) but has only recently attracted attention within the cognitive science modeling community (Kangasrääsiö, Jokinen, Oulasvirta, Howes, & Kaski, 2019; Turner & Van Zandt, 2014). ABC uses the discrepancy between observed data and model generated data for given parameter values to approximate the likelihood. For instance, suppose the model simulates data x^* for a parameter value θ^*; if the simulated data x^* is 'close' to the observed data y, then the parameter value θ^* would have higher likelihood. One approach to approximating the likelihood is to weight the proposal θ^* based on the difference between simulated and observed data, $D(x^*, y)$. The weights can be assigned using a Gaussian kernel density function, $\Psi(.|\delta)$, such that the difference $D(x^*, y)$ comes from a Gaussian distribution with mean zero and standard deviation σ. ABC uses summary statistics $S(\cdot)$ – such as the mean – of the observed and the simulated data to compute the difference, $D(x, y)$, such that $D(x, y) \approx D(S(x), S(y))$. So, the weighting function, $\Psi(D(x^*, y)|\delta)$ approximates the likelihood for a given parameter value.

We use ABC Sequential Monte Carlo (ABC-SMC) algorithm (Sisson, Fan, & Tanaka, 2007; Toni, Welch, Strelkowa, Ipsen, & Stumpf, 2009) to estimate the posterior distributions of the parameters in our models. The algorithm uses the idea of particle filtering: the posterior distribution evolves through successive populations of proposals, called particles; at each step, particles from the previous population are sampled, perturbed, evaluated, and filtered to form the next population. The steps of an ABC-SMC sampler are shown in Algorithm K1.

To determine the exact weighting function (Line 10 in Algorithm K1), we need to make reasonable choices about the tolerance parameter δ_t and summary statistics $S(.)$ depending on our data and the models. Our training data consist of estimates of number agreement effects in grammatical and ungrammatical conditions from 16 studies. Suppose i indexes ith study such that $i \leq 16$, and g and u index grammatical and ungrammatical data respectively; $y_{i,g}$ represents grammatical data from the ith study, and $y_{i,u}$ represents ungrammatical data from the ith study. A sample $\theta^{t,k}$, which is the kth particle in population t, is weighted as follows (Line 10 in Algorithm K1):

$$W_{t,k} = \frac{\left(\prod_{i=1}^{16} \Psi\left(D(S(x), S(y_{i,g})) \mid \delta_{t,i,g}\right) \cdot \Psi\left(D(S(x), S(y_{i,u})) \mid \delta_{t,i,u}\right)\right) \cdot \pi(\theta_{t,k})}{\sum_{j=1}^{K} W_{t-1,j} Q(\theta_{t-1,j}|\theta_{t,k})} \tag{K.2}$$

where $S(x)$ represents the summary statistic of the model-generated data; $S(y_{i,g})$ and $S(y_{i,u})$ represent the summary statistic of grammatical and ungrammatical data from study i. We choose one summary statistic, the *mean* of the data. This is because the posterior estimation accuracy does not really improve if we add more summary statistics other than the mean.

The term $\delta_{t,i,g}$ in the above equation represents the tolerance parameter associated with sampling population t and grammatical data of ith study, and $\delta_{t,i,u}$ represents the tolerance parameter associated with sampling population t and ungrammatical data of ith study. We determine $\delta_{t,i,g}$ and $\delta_{t,i,g}$ as follows:

$$\delta_{t,i,g} = \frac{3\sigma_{y_{i,g}}}{t} \; ; \; \delta_{t,i,u} = \frac{3\sigma_{y_{i,u}}}{t}$$

where $\sigma_{y_{i,g}}$ and $\sigma_{y_{i,u}}$ represent the standard deviation of grammatical and ungrammatical data from study i; t indexes the sampling population in the ABC-SMC sampler (see Algorithm K1). The parameter δ determines the degree of approximation of the likelihood: the smaller the value of δ, the better is the approximation. In ABC-SMC samplers, the tolerance parameter is (usually) adaptive to the successive populations, so that the approximation becomes better and better as the sampling proceeds. In our sampler, the value of δ is inversely proportional to the population index: the δ becomes smaller and smaller as the successive populations are sampled.

But why do we make δ adaptive to the data? The training data from 16 studies differs a lot in terms of uncertainty in the estimates of number agreement effects. This is because these studies had varying sample sizes: some studies had less than 50 participants while some had over 100 participants. We want the proposal weighting function to be sensitive to this information such that a dataset with larger sample size should contribute more in determining the weight of the proposed sample compared to a dataset with smaller sample size. To capture this information, we make the tolerance parameter adaptive to the variance of the training data. The value of δ in our sample is directly proportional to standard deviation of the data from study i; this function ensures that a dataset with smaller variance would have more impact on likelihood approximation. The term 3 in the numerator is a scaling factor to ensure that δ is large enough to sample from the high posterior density regions in a reasonable amount of time. If the δ becomes too small, most proposals would be weighted zero, and we will need a very large number of proposal, and consequently a huge amount of time, for getting a reasonable approximation of the posterior.

The term $Q(\theta_{t-1,j}|\theta_{t,k})$ in Eq. (K.2) is the backward transition kernel that determines the probability density of generating a sample $\theta_{t-1,j}$

Algorithm K1

ABC-SMC algorithm for parameter estimation: Given the observed data y, prior distribution $\pi(\theta)$, we have to estimate the posterior distribution of the parameter θ. The density function $Q(\theta|\theta^*)$ is the transition kernel; t indexes the successive populations and k indexes the particles in a population.

1 In population t=1
2 Initialize a population of K samples for the parameter θ as $\theta_{1,1:K}$
3 Set equal weights for each sample of θ as $W_{1,1:K} = 1/K$
4 **for population** $2 \leq t \leq T$
5 **for sample** $1 \leq k \leq K$
6 Sample θ^* from the previous population $\theta_{t-1,1:K}$ with weights $W_{t-1,1:K}$
7 Perturb θ^* by sampling $\theta^{**} \sim Q(\theta|\theta^*)$
8 Simulate data from the model given parameter value $\theta^{**} : x \sim Model(\theta^{**})$
9 Set $\theta_{t,k} = \theta^{**}$
10 Calculate weight for $\theta_{t,k}$ as $W_{t,k} = \frac{\Psi(D(S(x),S(y))|\delta_t)\cdot\pi(\theta_{t,k})}{\sum_{j=1}^{K} W_{t-1,j} Q(\theta_{t-1,j}|\theta_{t,k})}$
11 **end for**
12 Normalize the weights
13 **end for**

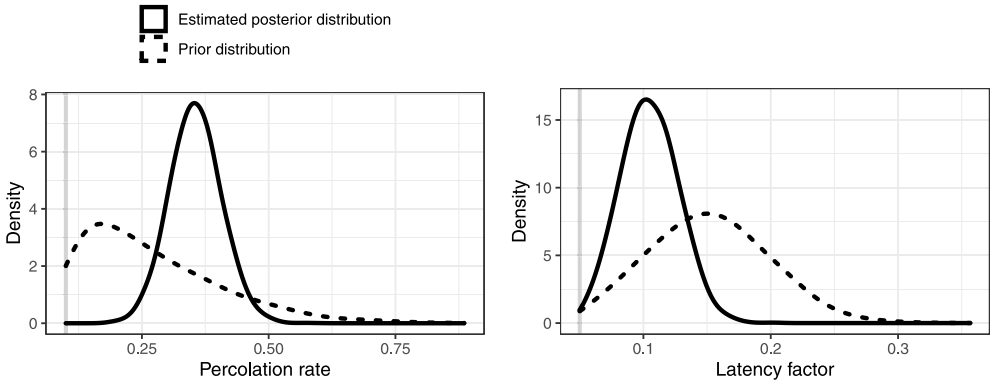


Fig. K.2. Estimated posteriors for the percolation rate and the latency factor parameter are shown using solid-lined density graphs; the corresponding prior distributions are shown using dash-lined density graphs.

in the previous population from the current sample $\theta_{t,k}$. The same transition kernel is also used in Line 7 in the Algorithm K1. The sample drawn from the previous population is perturbed using a transition kernel, $Q(\theta|\theta^*)$. Suppose θ^* is a sample from the previous population $t-1$; the sample θ^* is perturbed to get a new proposal θ^{**} such that $\theta^{**} \sim Q(\theta|\theta^*)$. We use a Gaussian distribution with standard deviation 0.02 for the transition kernel. The proposal θ^{**} is generated using $\theta^{**} \sim Normal(\theta^*, 0.02)$. The standard deviation of 0.02 is chosen such that the proposal θ^{**} does not lie too far away from θ^*, which is likely to be in the higher posterior density region, and also not too close to θ^* that sampler does not explore the parameter space beyond the previous population.

Fig. K.1 shows that ABC-SMC algorithm is able to accurately estimate the parameter values that were used to generate data. We validated the algorithm's performance as follows. Using a set of parameter values (latency factor 0.2, percolation rate 0.3), we generated fake data from the feature percolation-plus-retrieval model. Taking this fake data as the observed data, we used ABC-SMC algorithm to estimate model parameters. Fig. K.1 compares the estimated posterior distributions against the 'true' parameter values that were used to generate fake data.

Using the above ABC-SMC algorithm, we estimated parameters for the five models on 17 sets of training data. The parameter estimation time varied across models from 2 to 20 h. For example, the feature percolation-plus-retrieval model took approximately 3 h.[21] Fig. K.2 shows the estimated posterior distributions for the percolation rate and the latency factor parameter of the feature percolation-plus-retrieval model when the model was fitted to one of the training sets.

Cross-validation method

We compute predictive accuracy of a model using cross-validation (Stone, 1977; Vehtari et al., 2017).

As described in the following steps, our method is an extension of the k-fold cross-validation technique in which a dataset is partitioned into k subsets and each subset is iteratively held out for testing the model. Since we have 17 datasets here, we hold out each full dataset for testing.

(1) Prepare 17 sets of training and test data by leaving out one dataset as the test data and taking other 16 as training data
(2) In each iteration i (ranging from 1 to 17), fit the model on the training data, $y_{train,i}$ using approximate Bayesian computation as described in section 'Parameter estimation', and get the posterior distribution of the parameters of the model, say $\hat{\pi}(\Theta|y_{train,i})$; where $\hat{\pi}$ represent that it is an estimate of the true posterior $\pi(\Theta|y_{train,i})$.
(3) Compute the log predictive density of the model (explained below) in the ith iteration, i.e., lpd_i using corresponding test data, $y_{test,i}$
(4) After getting log predictive densities, lpd_i for 17 iterations, compute the expected log predictive density of the model, $\widehat{elpd}$ as the sum of all lpd_i

[21] The reproducible posterior estimation code for each model can be found at https://osf.io/gqj3p/

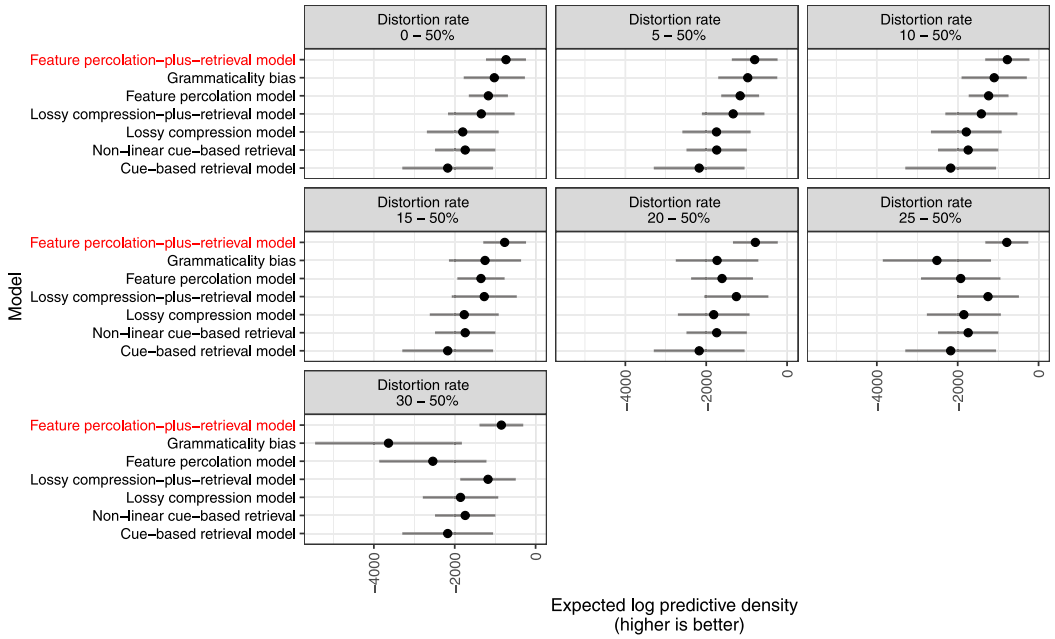


Fig. M.1. The predictive accuracies of the five models on 17 datasets under different assumptions about the distortion rate: Less-negative values imply better performance, and large-negative values imply poorer performance. The best-performing model is highlighted in red. The error bar around an $\widehat{elpd}$ value shows its standard error multiplied by two.

Let us unpack steps (3) and (4).

The log predictive density of a model in iteration i is given by logarithm of the average probability density of observing the test data $y_{test,i}$ given samples from the estimated posterior, $\hat{\pi}(\Theta|y_{train,i})$

$$lpd_i = \log \frac{1}{N} \sum_{j=1}^{N} \pi(y_{test,i}|\Theta_j) \text{ with } \Theta_j \sim \hat{\pi}(\Theta|y_{train,i}) \quad \text{(L.1)}$$

where j indexes samples from the posterior, N is the total number of samples drawn from the posterior, Θ_j represents the jth sample from the posterior. We apply step (3) in 17 iterations i.e., for 17 sets of training and test data and get an estimates of lpd_i in each iteration.

Because the likelihood for the most of our models is difficult to be expressed in explicit functional form (see section 'Parameter estimation'), we cannot directly estimate the likelihood of test data, $\pi(y_{test,i}|\Theta_j)$. We use a Gaussian kernel $\Psi(.|\delta)$, similar to approximate Bayesian computation in section 'Parameter estimation', to approximate the likelihood for each sample of parameter values

$$\pi(y_{test,i}|\Theta_j) \approx \Psi\big(D(S(y_{test,i}), S(y_{sim,j}))|\delta_i\big) \text{ with } y_{sim,j} \sim Model(\Theta_j) \quad \text{(L.2)}$$

where the statement $y_{sim,j} \sim Model(\Theta_j)$ means that the data $y_{sim,j}$ is simulated from the model conditioned on parameter value Θ_j; $D(S(y_{test,i}), S(y_{sim,j}))$ represents the difference between the test data and model generated data for the jth sample from the posterior. We use the mean summary statistic of the test and simulated data to calculate $D(S(y_{test,i}), S(y_{sim,j}))$ (see section 'Parameter estimation'). The kernel density function $\Psi\big(D(S(y_{test,i}), S(y_{sim,j}))|\delta_i\big)$ weights the parameter samples such that the samples that generate data close to the test data $y_{test,i}$ are weighted higher than those generate data far from $y_{test,i}$. The term δ_i determines the degree of approximation of the likelihood: as δ_i approaches zero, the approximation becomes exact. We set δ_i to be sensitive the variance of test data using $\delta_i = \frac{\sigma_{y_{test,i}}}{k}$, where $\sigma_{y_{test,i}}$ is the standard deviation of the test data $y_{test,i}$; the constant term in the denominator i.e., k (here $k = 9$) scales the value of δ such that approximated likelihood is not zero. Making δ adaptive to the variance of test data ensures that a dataset with smaller variance (i.e., larger sample size) has more weightage in determining the evidence for a model compared to dataset with larger variance.

The expected log predictive density of a model, $\widehat{elpd}$, is computed as,

$$\widehat{elpd} = \sum_{i=1}^{n} lpd_i \quad \text{(L.3)}$$

where n represent the total number of training and test sets, i.e., $n = 17$

The standard error of $\widehat{elpd}$ can be calculated as,

$$SE_{\widehat{elpd}} = \sqrt{n \cdot Var(lpd_i)} \quad \text{(L.4)}$$

where $Var(lpd_i)$ represents the variance of the lpd_i values

Prior sensitivity analysis

To verify whether each model's performance is sensitive to the prior on the distortion rate, we first specify seven different priors on the distortion rate parameter and then compare models under each prior assumption.

The prior on the distortion rate θ is given by

$$\theta \sim Normal_{lb=t}(0, 0.25) \quad \text{(M.1)}$$

where $lb = t$ represents the lower bound of t on distortion rate values such that $t \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$; it expresses different prior assumptions about the degree of feature distortion in the model.

Fig. M.1 shows a comparison of the models' predictive performance – in terms of expected log predictive density ($\widehat{elpd}$) values – under different prior assumptions about the distortion rate. A larger value of $\widehat{elpd}$, i.e., a less negative value of $\widehat{elpd}$, implies higher predictive performance. For the data considered here, we find that

1. The cue-based retrieval model has the lowest $\widehat{elpd}$ values meaning that it shows the worst predictive performance among all the models.

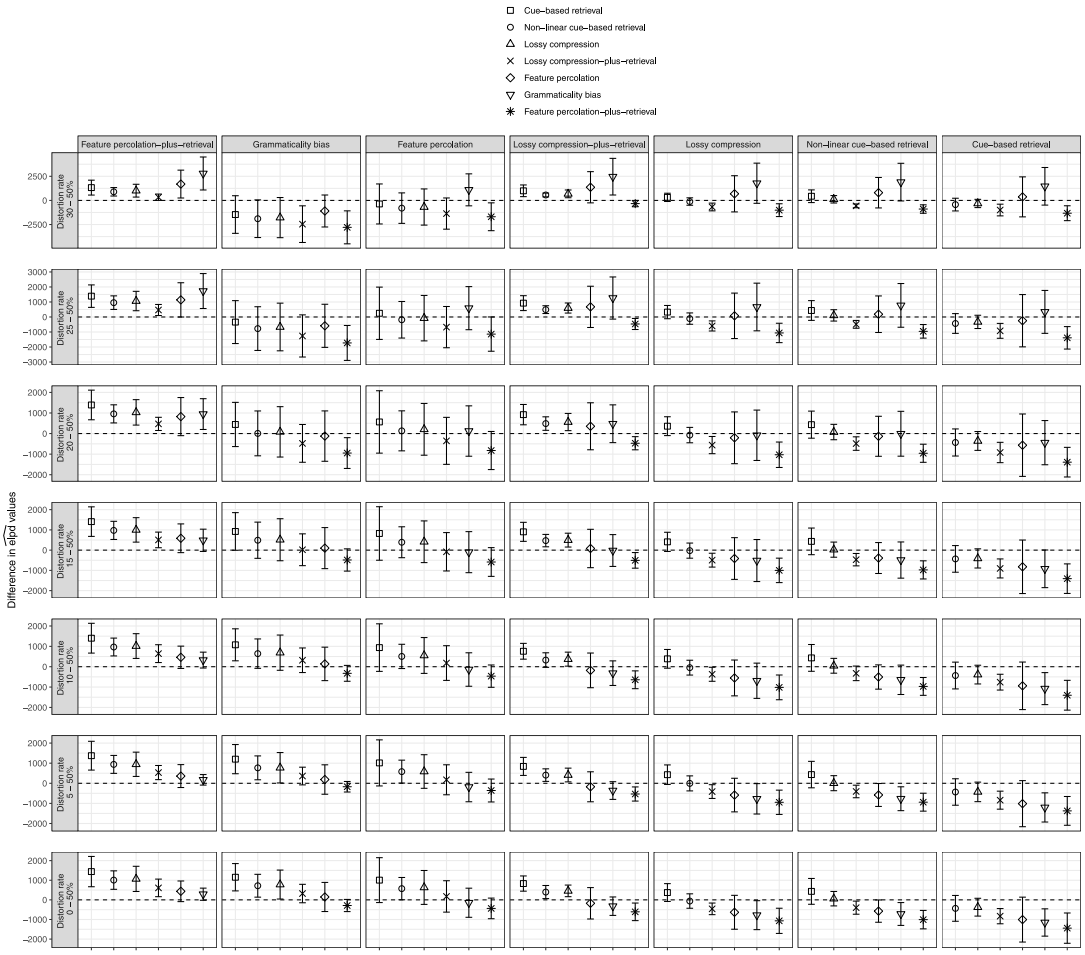


Fig. M.2. The difference in $\widehat{elpd}$ values of the models under different assumptions about the distortion rate: A positive difference in $\widehat{elpd}$ values means that the model shown in the facet's title performs better than the other models. Error bars show two times the standard error of the difference in $\widehat{elpd}$ values.

2. The new feature percolation-plus-retrieval model outperforms all other models under each prior assumption about the distortion rate.
3. The lossy compression-based models – the lossy compression model and the lossy compression-plus-retrieval model – perform better than the cue-based retrieval model but worse than the two feature percolation-based models. This pattern holds for most of the prior assumptions about the distortion rate, but when the distortion rate is greater than 25%, the lossy compression-plus-retrieval model performs better than the feature percolation model.

We also measure the difference in $\widehat{elpd}$ values ($\Delta\widehat{elpd}$) for each pair of models along with the standard error (SE) of difference. Fig. M.2 shows the $\Delta\widehat{elpd}$ values for each pair of models under seven different prior assumptions about the distortion rate. The positive difference in $\widehat{elpd}$ values implies that the model shown in a graph's title performs better than the other models.

The $\Delta\widehat{elpd}$ analysis reveals three key results. First, the feature percolation-plus-retrieval model shows positive $\Delta\widehat{elpd}$ values when compared with each of other four models. But the hybrid model is distinguishable from the feature percolation model only when the distortion rate is assumed to be greater than 25%. This implies that the feature percolation-plus-retrieval model outperforms the all other models except the feature percolation model, with which the hybrid model has comparable performance under most of the prior assumptions about distortion rate.

Second, the feature percolation model shows mostly positive $\Delta\widehat{elpd}$ values against the lossy compression model and the lossy compression-plus-retrieval model when the prior allows small distortion rates. But the error bars around the $\Delta\widehat{elpd}$ values always cross zero, implying that there is no clear evidence in the favor of feature percolation model over the two lossy compression-based models.

Third, the cue-based retrieval model shows negative $\Delta\widehat{elpd}$ values against the other four models, meaning that the cue-based retrieval model has the worst predictive performance of all the models considered. However, the error bars around the $\Delta\widehat{elpd}$ values indicate that the model's performance is not always distinguishable from the feature percolation model.

References

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485. http://dx.doi.org/10.1017/S0140525X00070801.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. http://dx.doi.org/10.1037/0033-295X.111.4.1036.

Anderson, J. R., & Lebiere, C. (2014). *The atomic components of thought*. Psychology Press, http://dx.doi.org/10.4324/9781315805696.

Anderson, J. R., et al. (1983). Retrieval of information from long-term memory. *Science*, *220*(4592), 25–30. http://dx.doi.org/10.1126/science.6828877.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, Article 104087. http://dx.doi.org/10.1016/j.jml.2020.104087.

Barker, J., Nicol, J., & Garrett, M. F. (2001). Semantic factors in the production of number agreement. *Journal of Psycholinguistic Research*, *30*(1), 91–114. http://dx.doi.org/10.1023/a:1005208308278.

Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. *Scientific Reports*, *6*(1), 1–14. http://dx.doi.org/10.1038/srep19203.

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7. http://dx.doi.org/10.1167/9.10.7.

Bock, K., & Cutting, C. J. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127.

Bock, K., & Eberhard, M. K. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, *8*(1), 57–99. http://dx.doi.org/10.1080/01690969308406949.

Bock, K., Eberhard, K. M., & Cutting, C. J. (2004). Producing number agreement: How pronouns equal verbsow pronouns equal verbs. *Journal of Memory and Language*, *51*, 251–278.

Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, *43*(2), 83–128.

Bock, K., & Miller, A. C. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93. http://dx.doi.org/10.1016/0010-0285(91)90003-7.

Brehm, L., & Bock, K. (2013). What counts in grammatical number agreement? *Cognition*, *128*(2), 149–169.

Clifton, C., Frazier, L., & Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivista Di Linguistica*, *11*(1), 11–39.

Cunnings, I., & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, *28*(1–2), 188–219. http://dx.doi.org/10.1080/01690965.2010.548391.

Cunnings, I., & Sturt, P. (2018). Retrieval interference and sentence interpretation. *Journal of Memory and Language*, *102*, 16–27. http://dx.doi.org/10.1016/j.jml.2018.05.001.

Dillon, B. W., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*, 85–103. http://dx.doi.org/10.1016/j.jml.2013.04.003.

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. In *Linguistic evidence: empirical, theoretical, and computational perspectives* (pp. 145–165).

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language*, *36*, 147–164. http://dx.doi.org/10.1006/jmla.1996.2484.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531–559. http://dx.doi.org/10.1037/0033-295x.112.3.531.

Engbert, R., Rabe, M. M., Schwetlick, L., Seelig, S. A., Reich, S., & Vasishth, S. (2022). Data assimilation in dynamical cognitive science. *Trends in Cognitive Sciences*, *26*, 99–102. http://dx.doi.org/10.1016/j.tics.2021.11.006.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association in retrieval processes: A computational account computational account. *Cognitive Science*, *43*(12), Article e12800. http://dx.doi.org/10.1111/cogs.12800.

Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15. http://dx.doi.org/10.1111/1467-8721.00158.

Fific, M. (2014). Double jeopardy in inferring cognitive processes. *Frontiers in Psychology*, *5*(1130), http://dx.doi.org/10.3389/fpsyg.2014.01130.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991. http://dx.doi.org/10.3758/s13423-012-0322-y.

Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction syntactic analysis of attraction. *Cognition*, *101*(1), 173–216. http://dx.doi.org/10.1016/j.cognition.2005.10.003.

Franck, J., Soare, G., Frauenfelder, U. H., & Rizzi, L. (2010). Object interference in subject–verb agreement: The role of intermediate traces of movement. *Journal of Memory and Language*, *62*, 166–182. http://dx.doi.org/10.1016/j.jml.2009.11.001.

Franck, J., Vigliocco, G., & Nicol, J. (2002a). Attraction in sentence production: The role of syntactic structure. *Language and Cognitive Processes*, *17*(4), 371–404.

Franck, J., Vigliocco, G., & Nicol, J. (2002b). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4), 371–404.

Frazier, L. (1979). *On comprehending sentences: syntactic parsing strategiesyntactic parsing strategies* (Ph.D. thesis), Amherst, MA: University of Massachusetts.

Frazier, L. (1987). Sentence processing: A tutorial review tutorial review. In M. Coltheart (Ed.), *The psychology of reading. Vol. 12* (pp. 559–586). Hillsdale, NJ: Erlbaum.

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*. Springer Science & Business Media.

Futrell, R. (2019). Information-theoretic locality properties of natural language. In *Proceedings of the first workshop on quantitative syntax (quasy, syntaxfest 2019)* (pp. 2–15). http://dx.doi.org/10.18653/v1/W19-7902.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), Article e12814. http://dx.doi.org/10.1111/cogs.12814.

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. http://dx.doi.org/10.1177/1745691614551642.

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(20), 8051–8056. http://dx.doi.org/10.1073/pnas.1216438110.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14*(3), 225–248. http://dx.doi.org/10.1080/016909699386293.

Gillund, G., & Shiffrin, M. R. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411–1423. http://dx.doi.org/10.1037/0278-7393.27.6.1411.

Hale, J. T. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the 2nd meeting of the north American chapter of the association for computational linguistics* (pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, *110*, 70–104. http://dx.doi.org/10.1016/j.cogpsych.2019.01.001.

Häussler, J. (2009). *The emergence of attraction errors during sentence comprehension* Unpublished doctoral dissertation, Germany: University of Konstanz.

Hofmeister, P., & Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology*, *5*, http://dx.doi.org/10.3389/fpsyg.2014.01237.

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., & Vasishth, S. (2015). Teasing apart retrieval and encoding interference in the processing of anaphors. *Frontiers in Psychology*, *6*(506), http://dx.doi.org/10.3389/fpsyg.2015.00506.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339. http://dx.doi.org/10.1016/j.jml.2017.01.004.

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study large-sample study. *Journal of Memory and Language*, (111), http://dx.doi.org/10.1016/j.jml.2019.104063.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions, Vols. 2, 289*. John Wiley & Sons.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *The Annual Review of Psychology*, http://dx.doi.org/10.1146/annurev.psych.59.103006.093615.

Kangasrääsiö, A., Jokinen, J. P. P., Oulasvirta, A., Howes, A., & Kaski, S. (2019). Parameter inference for computational cognitive models with approximate Bayesian computation. *Cognitive Science*, *43*(6), Article e12738. http://dx.doi.org/10.1111/cogs.12738.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. http://dx.doi.org/10.1016/j.tics.2017.11.006.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. http://dx.doi.org/10.3758/s13423-016-1221-4.

Kwon, N., & Sturt, P. (2016). Attraction effects in honorific agreement in Korean. *Frontiers in Psychology*, *7*(1302).

Lago, S., Acuña Fariña, C., & Meseguer, E. (2021). The reading signatures of agreement attraction. *Open Mind*, 1–22. http://dx.doi.org/10.1162/opmi_a_00047.

Lago, S., & Felser, C. (2018). Agreement attraction in native and non-native speakers of German. *Applied Psycholinguistics*, *39*(3), 619–647. http://dx.doi.org/10.1017/S0142716417000601.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149. http://dx.doi.org/10.1016/j.jml.2015.02.002.

Laurinavichyute, A. (2021). *Similarity-based interference and faulty encoding accounts of sentence processing* dissertation, University of Potsdam, https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/50966/file/laurinavichyute_diss.pdf.

Laurinavichyute, A., & von der Malsburg, T. (2022). Semantic attraction in sentence comprehension. *Cognitive Science*, *46*(2), Article e13086. http://dx.doi.org/10.1111/cogs.13086.

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

Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 234–243). Association for Computational Linguistics.

Lewis, R. L. (1993). *An architecturally-based theory of human sentence comprehension* Unpublished doctoral dissertation, Pittsburgh, PA: Carnegie Mellon University.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93–115. http://dx.doi.org/10.1007/BF01708421.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. http://dx.doi.org/10.1207/s15516709cog0000_25.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454. http://dx.doi.org/10.1016/j.tics.2006.08.007.

Linzen, T., & Dupoux, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependenciess to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. http://dx.doi.org/10.1162/tacl_a_00115.

Linzen, T., & Jaeger, F. T. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*, 1382–1411.

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., Van het Nederend, M. L., Burchert, F., et al. (2021). A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science*, *45*(4), Article e12956.

Logačev, P., & Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*(2), 266–298. http://dx.doi.org/10.1111/cogs.12228.

Mann, J. (1982). Atmosphere or red herring? *Journal of General Psychology*, *106*, 159–163.

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*, 879–906.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.

McElree, B. (2003). Accessing recent events. *Psychology of Learning and Motivation*, *46*, 155–200. http://dx.doi.org/10.1016/S0079-7421(06)46005-9.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*, 67–91. http://dx.doi.org/10.1016/S0749-596X(02)00515-6.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons, http://dx.doi.org/10.1002/0471721182.

Mertzen, D., Paape, D., Dillon, B., Engbert, R., & Vasishth, S. (2022). Syntactic and semantic interference in sentence comprehension: support from English and German eye-tracking data.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory and Cognition*, *18*(3), 251–269. http://dx.doi.org/10.3758/BF03213879.

Nicenboim, B., Roettger, T. B., & Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*, *70*, 39–55. http://dx.doi.org/10.1016/j.wocn.2018.06.001.

Nicenboim, B., Schad, D. J., & Vasishth, S. (2022). Introduction to bayesian data analysis for cognitive science. https://vasishth.github.io/bayescogsci/. Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series.

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34.

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, *42*, 1075–1100. http://dx.doi.org/10.1111/cogs.12589.

Nicol, J. L. (1995). Effects of clausal structure on subject-verb agreement errors. *Journal of Psycholinguistic Research*, *24*(6), 507–516.

Nicol, J., Forster, K. I., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, *36*, 569–587. http://dx.doi.org/10.1006/jmla.1996.2497.

Nivre, J., Abrams, M., et al. (2018). *Universal dependencies 2.3*. LINDAT/CLARIN Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2895.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*, 601–626. http://dx.doi.org/10.1016/j.jml.2006.08.009.

Open Science Collaboration, et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. http://dx.doi.org/10.1126/science.aac4716.

Paape, D., Avetisyan, S., Lago, S., & Vasishth, S. (2021). Modeling misretrieval and feature substitution in agreement attraction: A computational evaluation. *Cognitive Science*, *45*(8), Article e13019. http://dx.doi.org/10.1111/cogs.13019.

Palestro, J. J., Sederberg, P. B., Osth, A. F., Van Zandt, T., & Turner, B. M. (2018). *Likelihood-free methods for cognitive science*. Springer, http://dx.doi.org/10.1007/978-3-319-72425-6.

Parker, D. (2019). Cue combinatorics in memory retrieval for anaphora. *Cognitive Science*, *43*(3), Article e12715.

Patson, N. D., & Husband, M. E. (2016). Misinterpretations in agreement and agreement attraction. *Quarterly Journal of Experimental Psychology*, *69*(5), 950–971. http://dx.doi.org/10.1080/17470218.2014.992445.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*, 427–456. http://dx.doi.org/10.1006/jmla.1999.2653.

Raab, H. D. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*(5 Series II), 574–590. http://dx.doi.org/10.1111/j.2164-0947.1962.tb01433.x.

Raaijmakers, J. G., & Shiffrin, M. R. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93.

Rabe, M. M., Chandra, J., Krügel, A., Seelig, S. A., Vasishth, S., & Engbert, R. (2021). A Bayesian approach to dynamical modeling of eye-movement control in reading of normal, mirrored, and scrambled texts. *Psychological Review*, *28*, 803–823. http://dx.doi.org/10.1037/rev0000268.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N 400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. http://dx.doi.org/10.1038/s41562-018-0406-4.

Resnik, P. (1992). Left–corner parsing and psychological plausibility. In *Proceedings of COLING* (pp. 191–197). address not known. Available from http://citeseer.nj.nec.com/56865.html.

Ryu, S. H., & Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 61–71).

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Towards a principled Bayesian workflow: A tutorial for cognitive science. *Psychological Methods*, *26*, 103–126. http://dx.doi.org/10.1037/met0000275.

Schad, D. J., Nicenboim, B., Bürkner, P. C., Betancourt, M., & Vasishth, S. (2022). Workflow techniques for the robust use of Bayes factorsayes factors. *Psychological Methods*, http://dx.doi.org/10.1037/met0000472.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture architecture. In *Proceedings of the 3rd workshop on challenges in the management of large corpora* (pp. 28–34).

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *LREC* (pp. 486–493).

Schlueter, Z., Williams, A., & Lau, E. (2018). Exploring the abstractness of number retrieval cues in the computation of subject-verb agreement in comprehension. *Journal of Memory and Language*, *99*, 74–89. http://dx.doi.org/10.1016/j.jml.2017.10.002, https://www.sciencedirect.com/science/article/pii/S0749596X17300840.

Schneider, D. W., & Anderson, R. J. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology*, *64*(3), 127–160. http://dx.doi.org/10.1016/j.cogpsych.2011.11.001.

Scotti, P. S., Hong, Y., Golomb, J. D., & Leber, A. B. (2021). Statistical learning as a reference point for memory distortions: Swap and shift errors. *Attention, Perception, & Psychophysics*, *83*(4), 1652–1672. http://dx.doi.org/10.3758/s13414-020-02236-3.

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate bayesian computationayesian computation*. CRC Press.

Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *104*(6), 1760–1765. http://dx.doi.org/10.1073/pnas.0607208104.

Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject–verb number agreement. *Cognitive Science*, *42*, 1043–1074. http://dx.doi.org/10.1111/cogs.12591.

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, *124*, Article 101356. http://dx.doi.org/10.1016/j.cogpsych.2020.101356.

Smith, G., & Vasishth, S. (2022). A software toolkit for modeling human sentence parsing: An approach using continuous-time, discrete-state stochastic dynamical systems. https://psyarxiv.com/dtazq/.

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, *60*(2), 308–327. http://dx.doi.org/10.1016/j.jml.2008.11.002.

Staub, A. (2010a). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86.

Staub, A. (2010b). Response time distributional evidence for distinct varieties of number attraction. *Cognition*, *114*(3), 447–454. http://dx.doi.org/10.1016/j.cognition.2009.11.003.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterionkaike's criterion. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *39*(1), 44–47.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542–562.

Tanner, D. (2019). Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. *Cortex*, *111*, 210–237. http://dx.doi.org/10.1016/j.cortex.2018.10.011.

Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, *76*, 195–215.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, *6*(31), 187–202. http://dx.doi.org/10.1098/rsif.2008.0172.

Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from arabic. *Frontiers in Psychology*, *6*(347), http://dx.doi.org/10.3389/fpsyg.2015.00347.

Turner, B. M., & Van Zandt, T. (2014). Hierarchical approximate Bayesian computation. *Psychometrika*, *79*(2), 185–209. http://dx.doi.org/10.1007/s11336-013-9381-x.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 407–430. http://dx.doi.org/10.1037/0278-7393.33.2.407.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166. http://dx.doi.org/10.1016/j.jml.2006.03.007.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263. http://dx.doi.org/10.1016/j.jml.2011.05.002.

Vasishth, S. (2020). Using approximate Bayesian computation for estimating parameters in the cue-based retrieval model of sentence processing. *MethodsX*, *7*, Article 100850. http://dx.doi.org/10.1016/j.mex.2020.100850.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712.

Vasishth, S., & Engelmann, F. (2022). *Sentence comprehension as a cognitive process: a computational approach computational approach*. Cambridge, UK: Cambridge University Press, https://books.google.de/books?id=6KZKzgEACAAJ.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*, 1311–1342. http://dx.doi.org/10.1515/ling-2019-0051.

Vasishth, S., Jäger, L. A., & Nicenboim, B. (2017). Feature overwriting as a finite mixture process: evidence from comprehension data. In M. K. van Vugt, A. Banks, & W. Kennedy (Eds.), *Proceedings of the 15th international conference on cognitive modeling*. Coventry, UK: University of Warwick, https://arxiv.org/abs/1703.04081.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. http://dx.doi.org/10.1016/j.jml.2018.07.004, https://osf.io/eyphj.

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, *23*, 968–982. http://dx.doi.org/10.1016/j.tics.2019.09.003.

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from head-final structures. *Language and Cognitive Processes*, *25*(4), 533–567.

Vasishth, S., Yadav, H., Schad, D. J., & Nicenboim, B. (2022). Sample size determination for Bayesian hierarchical models commonly used in psycholinguistics. *Computational Brain & Behavior*, 1–25. http://dx.doi.org/10.1007/s42113-021-00125-y.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. http://dx.doi.org/10.1007/s11222-016-9696-4.

Vehtari, A., Ojanen, J., et al. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. http://dx.doi.org/10.1214/12-SS102.

Vigliocco, G., Butterworth, B., & Semenza, C. (1995). Constructing subject-verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, *34*, 186–215. http://dx.doi.org/10.1006/jmla.1995.1009.

Villata, S., & Franck, J. (2020). Similarity-based interference in agreement comprehension and production: Evidence from object agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(1), 170. http://dx.doi.org/10.1037/xlm0000718.

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology*, *9*(2), http://dx.doi.org/10.3389/fpsyg.2018.00002.

Wagers, M. (2008). *The structure of memory meets memory for structure in linguistic cognition*. College Park: University of Maryland.

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*, 206–237. http://dx.doi.org/10.1016/j.jml.2009.04.002.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, *108*(1), 40–55.

Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using approximate Bayesian computation. *Open Mind*, Provisionally accepted.

Chapter 4

Article II

Individual differences in cue weighting in sentence comprehension: An evaluation using Approximate Bayesian Computation

Himanshu Yadav, Dario Paape, Garrett Smith, Brian W. Dillon, and Shravan Vasishth

Open Mind, 6:147–168, 2022

DOI: https://doi.org/10.1162/opmi_a_00052
Code: https://osf.io/3na9q/

OPEN MIND

Discoveries in Cognitive Science

an open access journal

Citation: Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2022). Individual Differences in Cue Weighting in Sentence Comprehension: An Evaluation Using Approximate Bayesian Computation. *Open Mind: Discoveries in Cognitive Science*, *6*, 1–24. https://doi.org/10.1162/opmi_a_00052

DOI: https://doi.org/10.1162/opmi_a_00052

Supplemental Materials: https://doi.org/10.1162/opmi_a_00052; https://osf.io/3na9q/

Received: 2 April 2021
Accepted: 4 March 2022

Competing Interests: The authors declare no conflict of interest.

Corresponding Author:
Himanshu Yadav
hyadav@uni-potsdam.de

The MIT Press

REPORT

Individual Differences in Cue Weighting in Sentence Comprehension: An Evaluation Using Approximate Bayesian Computation

Himanshu Yadav[1], Dario Paape[1], Garrett Smith[1], Brian W. Dillon[2], and Shravan Vasishth[1]

[1]Department of Linguistics, University of Potsdam, Germany
[2]Department of Linguistics, University of Massachusetts, USA

Keywords: individual differences, interference effect, cue-based retrieval, approximate Bayesian computation, hierarchical modeling

ABSTRACT

Cue-based retrieval theories of sentence processing assume that syntactic dependencies are resolved through a content-addressable search process. An important recent claim is that in certain dependency types, the retrieval cues are weighted such that one cue dominates. This cue-weighting proposal aims to explain the observed average behavior, but here we show that there is systematic individual-level variation in cue weighting. Using the Lewis and Vasishth cue-based retrieval model, we estimated individual-level parameters for reading speed and cue weighting using 13 published datasets; hierarchical approximate Bayesian computation (ABC) was used to estimate the parameters. The modeling reveals a nuanced picture of cue weighting: we find support for the idea that some participants weight cues differentially, but not all participants do. Only fast readers tend to have the predicted higher weighting for structural cues, suggesting that reading proficiency (approximated here by reading speed) might be associated with cue weighting. A broader achievement of the work is to demonstrate how individual differences can be investigated in computational models of sentence processing without compromising the complexity of the model.

INTRODUCTION

A well-established claim in sentence processing is that dependency completion—establishing who did what to whom—is driven by a cue-based retrieval process (Lewis & Vasishth, 2005; McElree, 2000; Van Dyke, 2007; Van Dyke & McElree, 2011). The key idea behind cue-based retrieval is that codependents like verbs and their associated subjects are identified and connected together via a content-addressable search in memory.

For example, consider the ungrammatical sentences in (1) below.

(1) a. *The bodybuilder who worked with the trainers were complaining ...
b. *The bodybuilder who worked with the trainer were complaining ...

Within the cue-based retrieval framework, the auxiliary verb *were* in (1) is assumed to initiate a search in memory for an appropriate subject noun phrase, using feature specifications

such as plural (PL) and subject (SUBJ).[1] In example (1a), one of the retrieval cues (SUBJ) matches with the subject, *bodybuilder*, which is the correct codependent of *were*. The other retrieval cue (PL), matches a distractor noun, *trainers*. A signature property of the retrieval framework is that a distractor noun can be probabilistically misretrieved due to such a partial feature match (Paape et al., 2021; Vasishth et al., 2019). These misretrievals have the effect that reading time at the auxiliary *were* is faster in (1a) compared to (1b), in which the distractor does not match any of the retrieval cues. The explanation for this so-called *facilitatory interference* effect is a race process resulting in statistical facilitation (Raab, 1962): In every trial, two processes are initiated simultaneously, each corresponding to one matching feature (SUBJ: *bodybuilder*, PL: *trainers*). Whichever process finishes first is the winner in that trial. Statistical facilitation refers to the fact that when the mean finishing times of the two processes are very similar, the mean reading time resulting from the race process will be faster than both of the mean finishing times corresponding to the individual processes.[2] Facilitatory interference has been robustly observed in number-agreement configurations such as (1).

Similar effects have been shown in plausibility mismatch configurations (Cunnings & Sturt, 2018), negative polarity item licensing (Drenhaus et al., 2005; Vasishth et al., 2008; Xiang et al., 2009), and honorific processing (Kwon & Sturt, 2016). Given that cue-based retrieval parsing is intended as a comprehensive model of sentence processing, the phenomenon should generalize to any construction in which two partially matching nouns are available as retrieval candidates.

However, a construction that has been argued by Dillon et al. (2013) to be immune to facilitatory interference effects are antecedent-reflexive dependencies, as shown in the ungrammatical sentences in (2).

(2) a. *The bodybuilder who worked with the trainers injured themselves ...
b. *The bodybuilder who worked with the trainer injured themselves ...

Building on work by Sturt (2003), Dillon et al. (2013) argue that the search for an antecedent of a reflexive is guided exclusively by Principle A of the binding theory (Chomsky, 1981), which states that an anaphor must be bound by a c-commander within the same clause (see also Nicol & Swinney, 1989). The specific claim is that in examples like (2a) vs (2b), number marking on the reflexive (*himself* vs. *themselves*) is not used as a retrieval cue. This has the consequence that no difference in reading time is predicted at the reflexive in (2a) vs (2b). Other researchers (Cunnings & Sturt, 2014; Kush, 2013; Parker & Phillips, 2017) have proposed that although the c-command and plural retrieval cues are present at the reflexive, they are weighted differently in reflexives compared to subject-verb agreement: In reflexives, the weight of the structural cue versus the number cue is arguably higher, while in subject-verb agreement the cues have equal weights.[3] The increased cue weighting reduces or

[1] SUBJ is an abstract proxy feature that stands in for all kinds of information that may help identify subjects, including but not limited to a structural dominance relationship with the verb (c-command), case information, and linear order.

[2] Other explanations for the observed facilitatory interference effect have been proposed in the literature; in this article, we only investigate the implications of the cue-based retrieval account.

[3] There are some complications associated with treating c-command as a retrieval cue: c-command is a relation between two nodes in a tree, which must be determined dynamically as the syntactic tree is incrementally built. This issue is generally circumvented by the simplifying assumption that the antecedent is in the subject position of the sentence. See the discussion in Dillon et al. (2013).

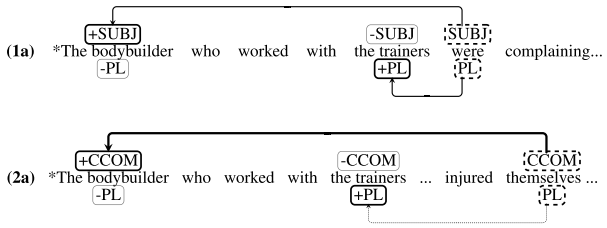


Figure 1. The cue-weighting proposal: A dashed box represents a retrieval cue, a thick box represents a feature that matches a retrieval cue, and a thin box represents a feature that does not match a retrieval cue. In agreement dependencies, the cues are weighted equally, while in reflexive dependencies the c-command cue is assumed to be weighted more highly than the number cue, as indicated by the thickness of the lines.

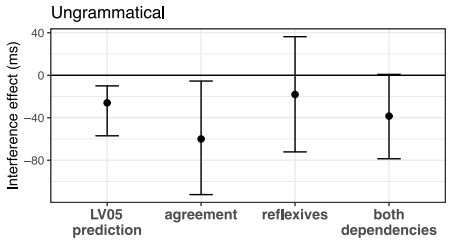


Figure 2. The facilitatory interference effect in ungrammatical agreement vs. reflexive dependencies in the Dillon et al. (2013) data; also shown is the cue-based retrieval model's predicted facilitatory interference effect for both dependency types (marked "LV05"). The figure is reused with permission from Jäger et al. (2020).

eliminates the role of the plural cue, so that very little or no facilitatory interference occurs. This is shown schematically in Figure 1.

Dillon et al. (2013) experimentally compared facilitatory interference effects in agreement and reflexive dependencies. They present data from 40 participants, which are summarized in Figure 2. Consistent with the cue-weighting account, we see a facilitatory interference effect in agreement dependencies, but no such effect in reflexive dependencies. The figure also shows the cue-based retrieval model's quantitative predictions with equal weights for the retrieval cues across dependencies.

Consistent with the standard practice in psycholinguistics, the claim made by Dillon et al. and others that structure has a privileged role at retrieval in reflexives focuses on average behavior across all items and participants. However, focusing only on average behavior misses two potentially important details.

First, the 40 individual participants in the Dillon et al. study show differences in both dependency types in the magnitude of the facilitatory interference effect. Figure 3 (top) shows that in both reflexives and agreement dependencies, the magnitude of the estimated effect for individuals ranges from 0 ms to effects as large as −150 ms.[4] This pattern of individual-level

[4] These estimates were extracted from a Bayesian hierarchical linear model fit to the data; the individual estimates, shown here with 80% and 95% Bayesian credible intervals, are so-called shrunken estimates for each participant. Such shrunken estimates are informed by the grand mean, and are more conservative than the estimates computed by estimating each individual's mean effect in isolation (Bates et al., 2014).

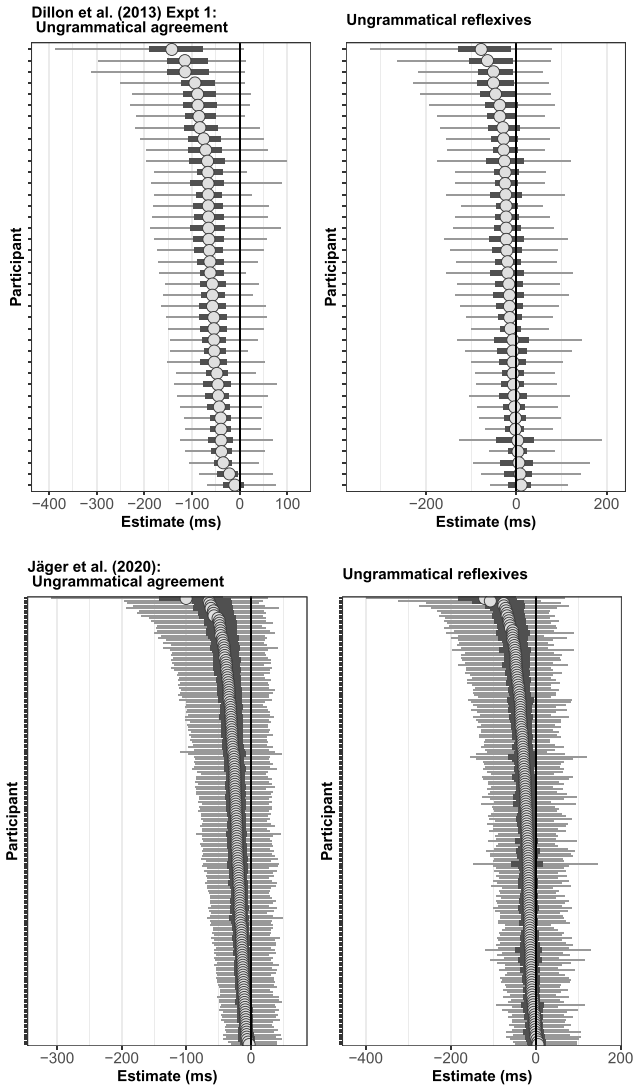


Figure 3. Individual-level facilitatory interference effects in ungrammatical agreement vs. reflexive dependencies in the Dillon et al. (2013) and Jäger et al. (2020) data. Shown are the shrunken estimates from a Bayesian hierarchical linear model fit to the data. The circle is the individual's mean effect, the solid error bar is an 80% Bayesian credible interval, and the thinner line a 95% credible interval.

variability is replicable: In a large-scale replication attempt of the Dillon et al. (2013) experiment, Jäger et al. (2020) found a similar pattern in 181 new participants; see Figure 3 (bottom). Thus, while the population-level estimates are consistent with the cue-weighting theory, some of the individual-level effects for reflexives across studies are not consistent with it: There are people who do show facilitatory interference for reflexives.

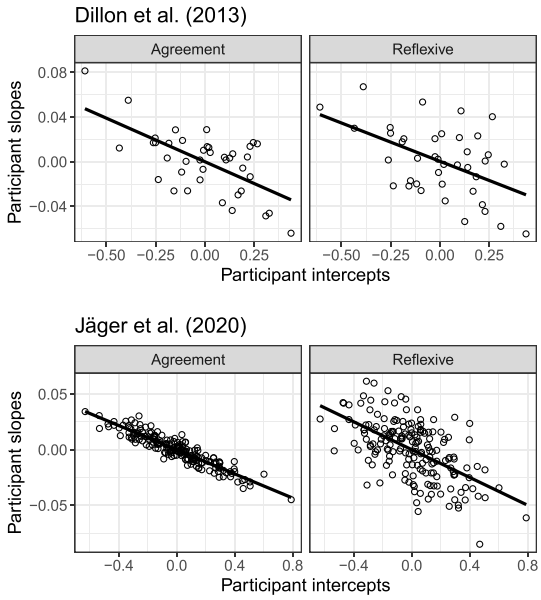


Figure 4. The relationship between mean reading times and the magnitude of the interference effect in the Dillon et al. (2013) data and the Jäger et al. (2020) data. Shown are the by-participant random intercept adjustments and the slope adjustments for the facilitatory interference effect. Both datasets show a clear pattern: slower participants show a larger magnitude of the facilitatory interference effect.

The second interesting aspect of the individual-level data is that the individual participants' average reading speed[5] is correlated with the magnitude of the facilitatory interference effect: the slower the participant, the larger the magnitude of the facilitatory interference effect. This correlation is displayed in Figure 4 for the Dillon et al. (2013) and the Jäger et al. (2020) datasets.

A simple explanation for this correlation would be that slower participants simply have higher standard deviations: a mean reading time with a relatively large value allows more room for variability around it than a mean reading time that has a small magnitude. There is, however, an alternative, theoretically grounded explanation for the observed variability as well as for the correlation with reading times: Individuals might differ in their cue weighting, and comparatively lower weight for the structural cue could be associated with slower reading speed, making less-fluent readers more susceptible to facilitatory interference.[6]

Under these assumptions, participants with higher cue weighting for the structural (c-command or subject) cue should show no facilitatory interference, while participants with approximately equal cue weighting should show facilitation.

[5] We use the term reading speed informally to refer to reading latency.

[6] Reading speed refers to an individual's overall reading fluency, which is assumed to be constant throughout the experimental trials. Since we do not have any independent measure of reading speed, as a first approximation we use reading times to estimate the reading speed parameter for an individual. We discuss this later in detail.

There is evidence for individual differences in cue weighting in the literature: memory research suggests that some individuals learn to use certain cues more effectively and reliably compared to other cues (Danker et al., 2011). Sohn et al. (2004) have shown that different training procedures lead participants to weight retrieval cues differently in a recall task. In sentence processing, some individuals could learn to use structural cues more effectively and reliably compared to nonstructural cues, and these individuals may also be more fluent readers. This view receives some support from the study reported by Traxler et al. (2012) on relative clauses: Traxler et al. found that readers with slower mean reading times experienced more processing difficulty in object relative clauses with misleading semantic cues (*The director that the movie pleased* ...) than readers with faster mean reading times. Traxler et al. hypothesized that due to fast readers' greater experience with the object-relative structure, the detrimental effect of the misleading semantic cue was reduced in fast readers.

Assuming that fast readers presumably also have more experience with other linguistic structures such as reflexives, their susceptibility to interference from structurally unavailable distractor nouns may be reduced in a similar manner.

Our proposed connection between language experience and cue weighting would also be consistent with results from nonnative speakers, whose reading speed has been found to correlate with their proficiency (Roberts & Felser, 2011), and who have been argued to assign higher weights to discourse-based cues over structural cues (Cunnings, 2017). In children, reading speed is positively correlated with text-level comprehension (Cutting & Scarborough, 2006; Jenkins et al., 2003), and among adults, highly skilled readers have been found to have shorter average fixation durations in eye-tracking than less-skilled readers (Underwood et al., 1990). Furthermore, reading speed as a measure has high reliability (Cunnings & Fujita, 2020; James et al., 2018), suggesting that it captures stable underlying differences between individuals.

Given the possibility of differences in cue weighting among individuals, the claim by Dillon et al. (2013) and others could be reformulated as follows: in reflexives, the majority of the individual participants should have higher cue weighting for the c-command cue than for the plurality cue. By contrast, in subject-verb number agreement, the majority of participants should exhibit equal cue weighting for the subject and plurality cues. Furthermore, if faster, more-skilled readers are less vulnerable to facilitatory interference, faster reading speed should correlate with higher weight for the structural cue over the plurality cue.

These predictions about individual-level variability in cue weighting and reading speed can be investigated quantitatively in the Lewis and Vasishth (2005) cue-based retrieval model. The predictions can be evaluated by (a) estimating, within the cue-based retrieval model, cue weighting and reading speed parameters separately for each individual participant, (b) deriving the predicted reading times from the model for each participant, and then (c) comparing these predicted reading times to the observed reading times for individuals. Stable alignment between the predicted and observed values would suggest that the model adequately captures the assumed underlying processes. Furthermore, given the observed correlation between reading speed and facilitatory interference in the data, a correlation between the model parameters that encode reading speed and cue weighting is expected.

We discuss the details of the cue-based retrieval model and the relevant parameters next.

The Lewis and Vasishth (2005) Model

The Lewis and Vasishth (2005) cue-based retrieval model, abbreviated below as LV05, is based on Adaptive Control of Thought—Rational (ACT-R) (Anderson & Lebiere, 2014) and assumes

that during retrieval, activation spreads to the memory chunks that match the retrieval cue. The more cues are matched by a chunk, the more activation it receives.

Simplifying slightly (cf. Engelmann et al., 2019), the total activation of a memory chunk i is given by

$$A_i = B_i + \sum_j W_j S_{ji} + \epsilon_i \tag{1}$$

where B_i is the baseline activation of the chunk i determined by past retrievals. A memory chunk i receives spreading activation from all matching cues j depending on the associative strength S_{ji} between the cue j and the chunk i, and the cue's weight W_j. The amount of activation spread determines the total activation of the memory chunks.

ϵ_i is Gaussian noise added to activation of the chunk i, such that $\epsilon_i \sim Normal(0, \sigma)$; σ represents the standard deviation of the normal. The fact that the activation values are noisy is crucial to the facilitatory interference effect: Even when two chunks receive the same amount of activation from the retrieval cues—such as *bodybuilder* and *trainers* in (1a)—the final activation values and the associated latencies can differ from trial to trial due to noise.

In a particular trial, the memory chunk that happens to have the highest activation is retrieved.[7] The time taken to complete the retrieval is determined by the activation level of the retrieved item: when activation is higher, the retrieval is faster. The retrieval time, RT_i, of a chunk is a negative exponential function of its activation at the time of retrieval:

$$RT_i = Fe^{(-fA_i)}. \tag{2}$$

F and f are two scaling parameters—the latency factor and the latency exponent, respectively. Latency factor and latency exponent reflect "surface-level" processing independent of the activation level of the chunks and, unlike the activation level, do not vary between trials. The latency factor represents the *general reading speed* of an individual and may, inter alia, include lexical access time, encoding time and motor response time.

In multiple-match scenarios, as in (1a), the activation distributions—and therefore the latency distributions—of the two candidate chunks have a larger overlap compared to (1b). The larger the overlap, the larger the observed speedup due to statistical facilitation. Since retrieval time of a chunk is a function of cue weights, the amount of facilitatory interference is modulated by the relative weights of the retrieval cues; see Figure 5. If each candidate chunk matches one retrieval cue and the cues have equal weights, the spreading activation for both chunks is the same. By contrast, if the structural cue has higher weight, the target chunk receives more activation, leading to reduced overlap and less facilitation. To understand the inner workings of the model in detail, see Lewis and Vasishth (2005) and Engelmann et al. (2019).

A convenient way to operationalize cue weighting is as the ratio of the weights of structural versus the nonstructural cue used in the retrieval (Engelmann et al., 2019). Cue weighting larger than 1 means more weight is given to the structural *subject* cue over the nonstructural *number* cue.[8]

[7] In ACT-R, a retrieval threshold parameter determines whether the activation is high enough for successful retrieval; when the activation falls below this threshold, retrieval failure occurs. This is how occasional retrieval failures are modeled in ACT-R.

[8] Given our prior specifications for the cue weighting parameter, a large cue weighting here is approximately equal to 4.

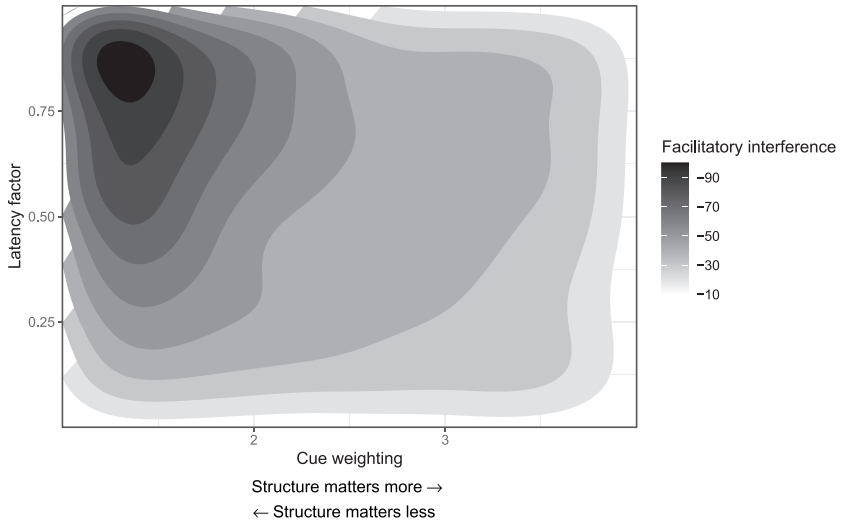


Figure 5. The facilitatory interference effect (in milliseconds) predicted by the model as a function of latency factor and cue weighting. The gradient from the light to darker shade represents the increasing magnitude of the facilitatory interference effect for the given values of cue weighting and latency factor. Cue weighting is the ratio of the weights of structural and nonstructural cues. The facilitatory effect increases linearly with latency factor when cue weighting is small; the effect decreases exponentially with increase in cue weighting.

For the present purposes, we are interested in estimating the parameters W_i, that is, the weights of the c-command and number cues, expressed as a ratio, and F, that is, the latency factor, for each individual participant. It is plausible to assume that the latency factor is the main source of variability in average reading speed between participants: The word-level processing of more experienced, fluent readers can be expected to be faster and more automatic (e.g., Joo et al., 2021; Kuperman & Van Dyke, 2011; Logan, 1997; Samuels & Flor, 1997), which should lead to faster reading irrespective of the memory processes triggered during the completion of dependencies. For the present purpose, we will use reading speed as a proxy for language experience.[9]

Our hypothesis is that fast word-level processing as indexed by small values of F tends to coincide with high values for the W_i parameter that controls the weight of structural over nonstructural cues used at the sentence level, and that language experience is the connecting factor. Participants who read fluently and automatically should therefore be less susceptible to interference from structurally illicit distractors, and there should thus be a negative correlation between latency factor and cue weighting.

Robust Estimation of Individual Differences Using Approximate Bayesian Computation

Parameter estimation is a key challenge in modeling individual differences. Suppose that we collect some data, y, from participants in an experiment, and we assume that the data y come from a function of a vector of parameters Θ: $y \sim f(\Theta)$. The statement $y \sim f(\Theta)$ represents a model where $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$. The aim of parameter estimation is to compute the join posterior distributions of the parameters Θ that would have generated the observed data y.

[9] This is a simplifying assumption, as reading speed may be influenced by a multitude of factors. We will discuss alternative approaches in the general discussion.

Parameter estimation for the LV05 model is difficult because the model is nondeterministic and its likelihood cannot easily be expressed analytically, unless we drastically simplify the model (e.g., in Lissón et al., 2021; Nicenboim & Vasishth, 2018; Paape et al., 2020). Therefore, we cannot employ the standard Bayesian estimation method for estimating parameters of the LV05 model.

In the published modeling work on cue-based retrieval, grid search is the standard approach used to estimate the parameters in the model (Engelmann et al., 2019; Mätzig et al., 2018). In grid search, the parameter space for each parameter θ_i is divided into n equally spaced points, $\{v_{i,1}, ..., v_{i,n}\}$ and all possible combinations are taken to create n^m grid points. Each grid point is evaluated by generating a model prediction, and a grid point that provides the best fit to the observed data is chosen as the estimated parameter value (or set of values, in case multiple parameter values provide the best fit).

There are several important limitations to the grid search method (Kangasrääsiö et al., 2019). First, grid search is a brute-force method, and therefore inefficient; the number of computations increases exponentially as the number of parameters m increases. It is therefore not computationally feasible to compute individual-level as well as population-level estimates using this approach. Second, grid search usually delivers a point estimate of the parameter value; uncertainty about the parameter estimates cannot be computed.

As Kangasrääsiö et al. (2019) point out, a better way to estimate model parameters is Bayesian estimation. This approach allows us to estimate the posterior distributions of the parameter values given the observed data, that is, $\pi(\theta|y)$ using Bayes's rule, which is shown in Equation 3.

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \tag{3}$$

As Bayes's rule states, the estimation of the posterior distribution, $\pi(\theta|y)$ requires knowledge about the likelihood, $\pi(y|\theta)$ (i.e., probability density of seeing the data for given parameter value) and the prior knowledge about θ before y is available, that is, $\pi(\theta)$. However, for the LV05 model, the likelihood function, that is, $\pi(y|\theta)$ is difficult to express mathematically.

For such situations, approximate Bayesian computation (ABC) has emerged as an effective tool for approximating the posterior distributions of the parameters (Palestro et al., 2018; Sisson et al., 2018). ABC uses an approximation of the likelihood function to estimate the posterior distribution for the parameters. Even when the likelihood function is unknown, the model can still generate simulated data for given parameter values. The ABC method consists of comparing the observed data with the simulated data. Since ABC is rooted in Bayesian estimation techniques, it also computes the uncertainty bounds on the parameter estimates.

A simple ABC sampler works as follows:

1. Draw a proposal θ^* from the prior distribution, that is, $\theta^* \sim \pi(\theta)$.
2. Simulate data x from the model conditional on θ^*.
3. If the simulated data x is "close" to the observed data y, accept θ^* as a sample for the posterior distribution. The "closeness" between simulated data x and observed data y, that is, $dist(x, y)$ approximates the likelihood for θ^*.

If the above steps are carried out repeatedly, we can obtain samples from the approximate posterior distribution.

One approach to approximating the likelihood is to weight the proposal θ^* based on distance between simulated and observed data, $dist(x, y)$. If the simulated data x corresponding to

a proposed value θ^* is closer to the observed data y, the proposed θ^* will have higher weight. The weights can be assigned using a distribution centered at zero, such that if $dist(x, y)$ is zero, the proposed θ^* will have highest weight, and as $dist(x, y)$ increases, weight decreases. ABC uses summary statistics $S(\cdot)$—such as the mean—of the observed and the simulated data to compute $dist(x, y)$. The weight assigning function is called a kernel function. Suppose that the kernel function is $\Psi(dist(x, y)|\delta)$.

Here, δ is the tolerance parameter, which determines the degree of the approximation. The lower the value of δ, the better the approximation. When δ approaches 0, the approximation becomes exact:

$$\pi(\theta|y) \propto \int_X \Psi(dist(S(x), S(y))|\delta)\pi(x|\theta)\pi(\theta)dx, \tag{4}$$

where X is the support of the simulated data from the model.

The intractable likelihood term $\pi(x|\theta)$ cancels out while calculating the probability of accepting a proposal in posterior simulation algorithms.

We use the ABC method to estimate the participant-level and population-level parameters of the cue-based retrieval model. Our parameter estimation problem is more complex than the simple ABC algorithm presented above: in order to model individual differences, we need to estimate both individual- and population-level parameters simultaneously. For such a situation, Turner and Van Zandt (2014) propose a hierarchical Gibbs ABC algorithm. The details of this method are explained later, but in essence, the idea is to first draw samples for the individual-level parameters using ABC and then sample for each population-level parameter from a distribution conditioned on all other parameters.

Returning to our main research question, we now describe a computational evaluation of the two claims that follow from the cue-weighting proposal: (i) in reflexives, most individuals should have higher cue weighting for the c-command cue compared to the plurality cue and (ii) in agreement, most individuals should have equal cue weighting for the subject and plurality cues. Furthermore, we investigate (iii) whether there is a correlation between cue weighting and reading speed.

MODELING INDIVIDUAL DIFFERENCES IN THE CUE-BASED RETRIEVAL MODEL

We obtained 13 datasets from four published studies (Dillon et al., 2013; Jäger et al., 2020; Lago et al., 2015; Wagers et al., 2009) that report the facilitatory interference effect. Table 1 lists the datasets along with number of participants in the experiment and population-level mean facilitatory interference effect. Out of these 13 datasets, 11 tested subject-verb agreement dependencies (Dillon et al., 2013; Lago et al., 2015; Wagers et al., 2009), and the remaining two datasets (Dillon et al., 2013; Jäger et al., 2020) investigated both subject-verb agreement and reflexive dependencies.

Using these datasets, we implemented hierarchical ABC to estimate the participant-level and population-level parameters. The code and data are available from https://osf.io/3na9q/. The latency factor (which modulates retrieval time) and cue weighting were estimated simultaneously for each participant; the computational details behind the ABC algorithm are explained in Appendix S1 in the Supplemental Materials.

For each dataset, we fit a Bayesian hierarchical model with varying intercepts and slopes for participants and items, where reading time is the dependent variable and condition (multiple-match vs. single-match) is a sum-coded predictor (Schad et al., 2020). As discussed earlier,

Table 1. Facilitatory interference effect data used for estimating parameters of the LV05 model.

Dataset	Number of Participants	Interference Effect (in milliseconds)
Subject-verb agreement dependency		
Dillon et al. (2013) Exp 1	40	−60 [−111, −11]
Jäger et al. (2020)	181	−27 [−47, 2]
Lago et al. (2015) Exp 1	32	−27 [−55, 0]
Lago et al. (2015) Exp 2	32	−23 [−54, 7]
Lago et al. (2015) Exp 3a	32	−13 [−29, 2]
Lago et al. (2015) Exp 3b	32	−13 [−33, 6]
Wagers et al. (2009) Exp 2	28	−23 [−49, 1]
Wagers et al. (2009) Exp 3 (Singular subject)	60	−18 [−46, 9]
Wagers et al. (2009) Exp 3 (Plural subject)	60	−4 [−39, 30]
Wagers et al. (2009) Exp 4	44	−28 [−53, −3]
Wagers et al. (2009) Exp 5	60	−20 [−44, −4]
Antecedent-reflexive dependency		
Dillon et al. (2013) Exp 1	40	−18 [−73, 36]
Jäger et al. (2020)	181	−23 [−49, 2]

Note. The table lists the published datasets along with the number of participants and population-level mean interference effect. The square brackets show 95% credible intervals around the population-level interference effects, that is, there is a 95% probability that the value of the population-level interference effect lies within this range, given the statistical model and data.

each fitted model provided shrunken estimates of individual-level facilitatory interference effects. We used these individual-level shrunken estimates as data for the cue-based retrieval model.

As we describe now, the hierarchical ABC method provided estimates of participant-level latency factor and cue weighting, population-level mean latency factor and cue weighting, population standard deviation for latency factor and cue weighting, and the correlation between latency factor and cue weighting.

Suppose that y_j are the data (interference effect) associated with the j^{th} participant. LF_j and CW_j are the latency factor and the cue weighting parameters, respectively, for the j^{th} participant. We assume that the human data y_j are generated by the LV05 model with the parameters LF_j and CW_j:

$$y_j \sim Model\left(LF_j, CW_j\right). \tag{5}$$

We further assume that LF_j and CW_j come from a bivariate normal distribution with population means μ_{LF} and μ_{CW}, standard deviations σ_{LF} and σ_{CW}, and correlation parameter, ρ:

$$\begin{pmatrix} LF_j \\ CW_j \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_{LF} \\ \mu_{CW} \end{pmatrix}, \begin{pmatrix} \sigma_{LF}^2 & \rho\sigma_{LF}\sigma_{CW} \\ \rho\sigma_{LF}\sigma_{CW} & \sigma_{CW}^2 \end{pmatrix}\right). \tag{6}$$

The goal here is to obtain posterior estimates of the individual-level parameters LF_j and CW_j and the population-level parameters μ_{LF}, μ_{CW}, σ_{LF}, σ_{CW}, and ρ. Assuming that the data come from n participants, we have $2n + 5$ parameters to estimate, that is, LF_j, CW_j for each participant and five population-level parameters μ_{LF}, μ_{CW}, σ_{LF}, σ_{CW}, and ρ.

Results

Cue Weighting in Agreement and Reflexive Dependencies. First, we focus on the Dillon et al. (2013) and Jäger et al. (2020) experiments. Figures 6 and 7 compare the distribution of participant-level cue weighting in agreement and reflexive dependencies for these two studies.

The individual-level estimates for agreement show that in the Dillon et al. (2013) and Jäger et al. (2020) data, almost all participants have equal cue weighting, that is, the weight ratio between the structural cue and the number cue is close to 1:1, consistent with the cue-weighting account. For reflexives, the two experiments show similar patterns: About one quarter of participants in the original Dillon et al. (2013) have estimates for cue weighting close to or higher than 2:1 in favor of the structural cue, which is consistent with the cue-weighting

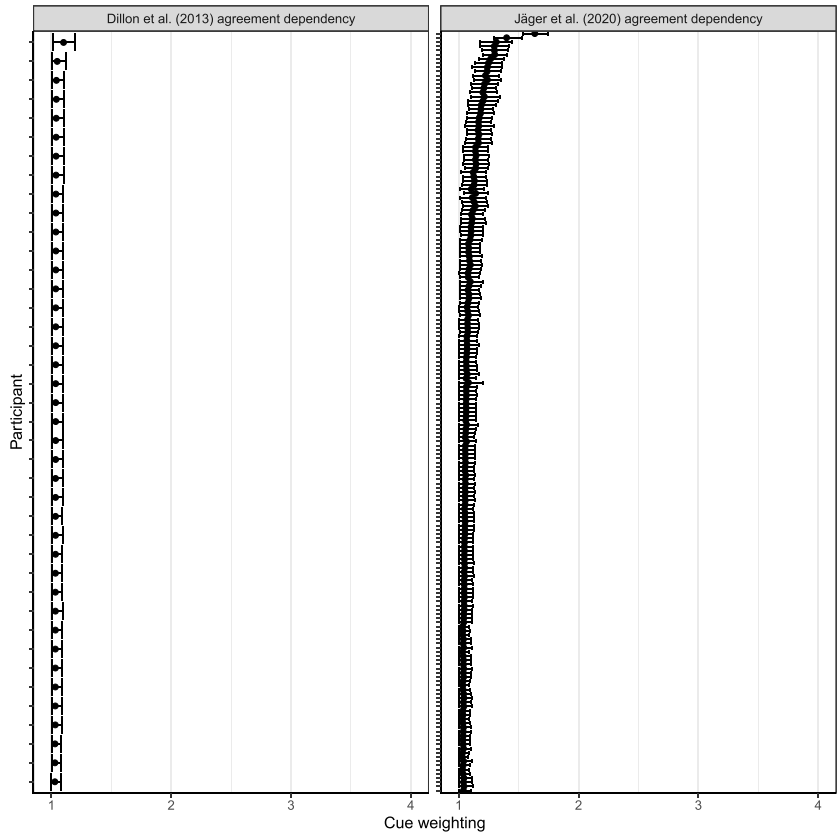


Figure 6. Participant-level cue weighting for agreement dependency data in Dillon et al. (2013) and Jäger et al. (2020). Shown are the 95% credible intervals for each individual's estimate, along with the estimated mean parameter value for each participant.

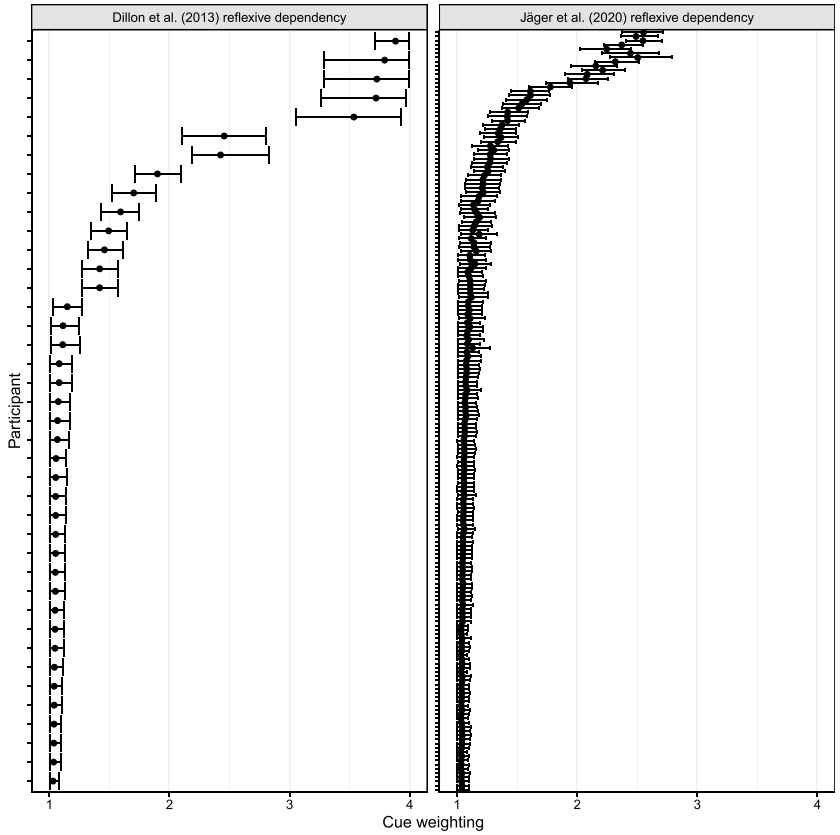


Figure 7. Participant-level cue weighting for reflexive dependency data in Dillon et al. (2013) and Jäger et al. (2020). Shown are the 95% credible intervals for each individual's estimate, along with the estimated mean parameter value for each participant.

account. At the same time, many participants have cue weighting close to 1:1. The reflexives data from Jäger et al.'s (2020) replication study also have a minority of participants with high cue weighting, but there are also many participants with cue weighting estimates close to 1:1. The overall pattern in the replication study is noticeably more graded, but this may be due to the larger sample size.

For the Jäger et al. (2020) study, simulating data based on the individual participants' parameter estimates for cue weighting and latency factor and computing the facilitatory interference effect yields values very close to the shrunken estimates based on the original data. Figure 8 shows the model estimates alongside the shrunken estimates from the hierarchical linear models. Although not shown here, the model can also capture the individual-level estimates from the Dillon et al. (2013) data (see Appendix).

Correlation Between Cue Weighting and Reading Speed. We now analyze each of the 13 datasets separately to investigate the relationship between cue weighting and latency factor. Figure 9 shows the posterior distribution of the correlation between cue weighting and reading speed, that is, the latency factor, across the 13 datasets. Most of the estimates are negative, in line with

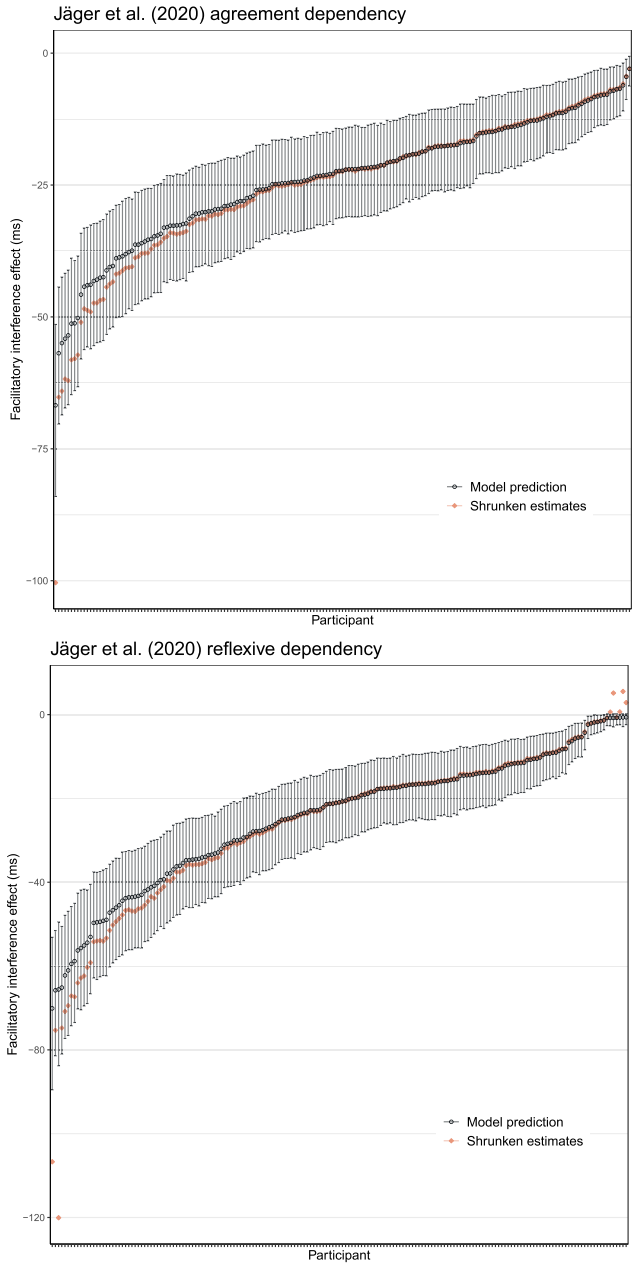


Figure 8. Posterior predicted interference effects for individuals in agreement and reflexive dependencies, derived from the cue-based retrieval model after estimating individual-level parameters for cue weighting. Shown are the posterior mean and 95% credible interval, along with the estimate of the shrunken mean for each participant.

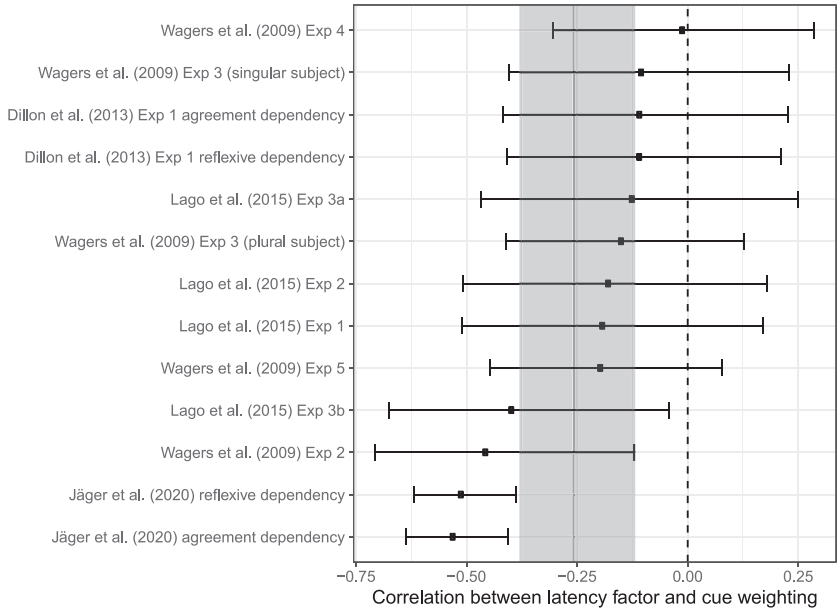


Figure 9. Estimated correlations between participant-level latency factor and cue weighting for 13 datasets that investigated the facilitatory interference effect. Shown are the posterior distributions of the correlation parameters (mean and 95% credible interval); the dark band is the posterior distribution of the overall correlation, derived from a random-effects meta-analysis.

the prediction that slower, less-skilled readers who have high latency factors should have low cue weighting, that is, weight ratios close to 1:1. In order to synthesize the evidence relating to the correlation from all the available studies, we conducted a random-effects meta-analysis. We use a Bayesian meta-analysis method that is based on the Fisher z-transformation of correlation estimates (Fisher, 1921; Zhang et al., 2017). The meta-analysis model is described in Appendix S2 in the Supplemental Materials.

The meta-analytical estimate for the correlation parameter is −0.26, and the associated 95% credible interval is [−0.38, −0.12], as shown by the shaded region in Figure 9. Taken together, the data thus indicate a negative correlation between cue weighting and latency factor, consistent with the idea that slower readers tend to have equal cue weighting, and that the faster the participant, the higher the weighting in favor of the structural cue.

Discussion

Based on the two studies that investigated both subject-verb agreement and reflexive dependencies, Dillon et al. (2013) and Jäger et al. (2020), we compared individual-level cue weighting for each of the dependencies. Results for agreement dependencies showed equal cue weighting for the majority of participants in both experiments, consistent with the cue-weighting account. By contrast, the results for reflexive dependencies showed that both the Dillon et al. (2013) and Jäger et al. (2020) data do have some participants with higher cue weighting for the structural cue, but the majority of the participants do not have higher cue weighting.

It is interesting to note here that, based on the average estimates from their respective datasets, Dillon et al. (2013) and Jäger et al. (2020) came to different conclusions. Dillon et al. concluded that reflexives are processed differently from subject-verb agreement, in that number features are effectively ignored during retrieval due to Principle A. In contrast, Jäger et al. concluded, again based on average estimates, that their larger dataset had showed no indication of a difference in processing between agreement and reflexive dependencies.

The individual-level estimates qualify both conclusions: in both the Dillon et al. (2013) and the Jäger et al. (2020) data, there are in fact participants that mostly ignore the number features of structurally inaccessible noun phrases in reflexive dependencies, as indicated by higher weighting of the structural cue. On the other hand, a majority of participants show processing profiles for reflexives that are similar to agreement dependencies, where the number cue of the distractor matters.

The present findings are thus inconsistent with the strongest possible version of the cue-weighting proposal, which is that *all* readers should show high cue weighting for reflexives. They are also inconsistent with the next-strongest version of the proposal, which is that *most* readers should show high cue weighting for reflexives. Instead, our findings show that there is only a nonnegligible minority group of readers in the two studies who show high cue weighting for reflexives.

The meta-analysis of our simulations based on all 13 available datasets shows that cue weighting in favor of the structural cue correlates with reading speed, such that faster readers are more likely to assign higher weight to the structural cue. The correlation estimates vary between studies, and the credible intervals of the correlation parameter cross zero in many cases, especially for studies with much fewer participants than the higher-powered study of Jäger et al. (2020). Nevertheless, the evidence as a whole, as quantified by the meta-analysis, is compatible with the notion that faster, more skilled readers assign more weight to structural cues during sentence processing. While most of the studies used in our simulation tested subject-verb agreement dependencies, this relationship also holds for reflexive dependencies, as shown by the estimate for the Jäger et al. and Dillon et al. (2013) studies (see Figure 9).

GENERAL DISCUSSION

The present article addressed two questions related to individual differences in sentence comprehension. The first question was whether retrieval cues are weighted differently in subject-verb agreement and reflexive agreement configurations. The second question was whether participants who read more quickly on average also show larger facilitatory interference effects. In order to answer these questions, we used approximate Bayesian computation to fit individual- and population-level parameters of the cue-based retrieval model of Lewis and Vasishth (2005) to 13 datasets. ABC allows us to use well-motivated but complex models of sentence processing to directly test our questions using experimental data. We now discuss the results for the cue weighting and correlation questions in turn.

Cue Weighting in Agreement and Reflexive Dependencies

For the studies that provided data on reflexive dependencies (Dillon et al., 2013; Jäger et al., 2020), we found that a minority of participants did have higher cue weighting for the structural cue than the nonstructural cue in these dependencies, but also that most participants did not. This suggests that some participants do indeed strongly adhere to Principle A during online sentence comprehension. This prevents misretrievals of a distractor word and thus blocks the facilitatory interference effect that is often observed in number agreement contexts.

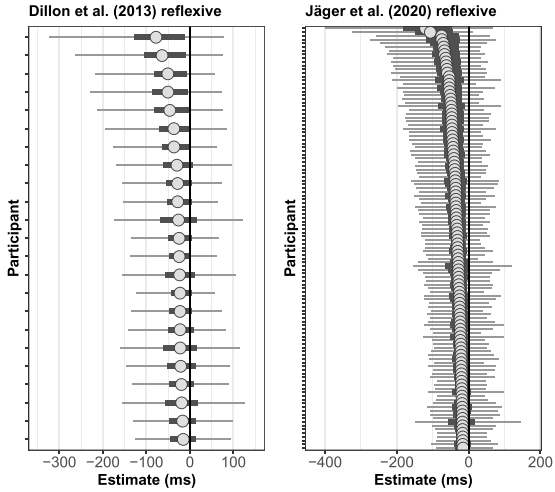


Figure 10. The facilitatory interference effect associated with the participants who had approximately equal cue weights for c-command and number cue in reflexive dependencies.

However, for the majority of participants, there is no indication of a difference in processing between agreement and reflexives. For participants who weigh structural and nonstructural cues approximately equally, facilitatory interference occurs in both constructions (see Figure 10). Overall, the results suggest that the cue-weighting hypothesis only holds for a subset of English native speakers.

The subset of participants who have higher weighting for the structural cue show weak or absent facilitatory interference, that is, a magnitude of smaller than 8 ms. The estimated value of cue weighting for these participants has a logarithmic relationship with facilitatory interference: the smaller the magnitude of the facilitatory interference effect, the higher the weighting for the structural cue over the nonstructural cue. By contrast, participants with facilitatory interference effects of a magnitude over 8 ms have equal cue weighting for structural and nonstructural cues.

For the subset of participants who do weight the structural cue more highly, it is reasonable to ask what the cause of the different weighting is. Two possible reasons are as follows:

1. There are differences in the "level of representation" accessed by each dependency (Dillon et al., 2013). The reflexive dependencies are possibly keyed to more semantic/notional number than are agreement dependencies, which causes reflexives to be less sensitive to morphosyntactic number and hence to have higher weight for the structural cue. This claim has independent support in Kreiner et al. (2013), who found that in processing reflexive dependencies where the antecedent and reflexive matched in notional number, the mismatch in morphosyntactic number between the reflexive and the antecedent noun did not incur a processing cost. By contrast, subject-verb agreement showed a reliable mismatch cost due to morphosyntactic number even when the subject noun and the verb had the same notional number.
2. Reflexives and agreement dependencies differ in predictability. In antecedent-reflexives constructions, the reflexive is generally not expected given the preceding context.

Hence, unlike agreement dependencies, there is no prediction for an upcoming co-dependent that could facilitate dependency completion. A possible implication is that comprehenders adopt a more conservative approach to resolving reflexive dependencies. This could change the priority given to different cues in the retrieval strategy used to resolve a reflexive, by giving more priority to the more diagnostic structural cues (see also Parker & Phillips, 2017).

However, the above possibilities are underspecified and untested in the context of the cue-based retrieval model, which we used in this study. In order to verify them, future work with cue-based retrieval theories will need to formalize how linguistic cues get learned over time for different dependencies.

A broader conclusion to be drawn from our simulations is that a focus on modeling population-level effects may mask theoretically interesting variation at the individual level. While it has become the norm in psycholinguistics to include random intercepts and slopes by participant and by item to guard against anticonservative conclusions at the population level (Barr et al., 2013), the magnitude and shape of the observed interindividual variation is seldom discussed. In principle, any two linguistic constructions may be distinguished both by the amount of possible variation between individuals, as has been claimed for subject-verb agreement versus reflexive dependencies, but also by whether differences between speakers are more quantitative or more qualitative in nature (Navarro et al., 2006; Rouder & Haaf, 2021). With regard to the weighting of structural cues in sentence processing, our results suggest that there is more variability between speakers for reflexive dependencies than for agreement dependencies. However, whether it is possible to experimentally identify a clearly delimitable subgroup of speakers who show strong adherence to Principle A during the processing of reflexive dependencies is not yet clear.

What underlying factors could plausibly distinguish readers with high cue weighting from readers with low cue weighting? One candidate factor we suggest is language experience, as indexed by reading speed (see discussion below). Nevertheless, it is an open question whether readers with a high weighting of the structural cue generally have a more strongly developed or less "gradient" grammar compared to those with low cue weighting. Rather than having a general preference for structural cues, it is also possible that they treat Principle A in particular as a "hard" constraint, whereas other participants may treat it as more of a "soft" constraint (Sorace & Keller, 2005). This latter view seems to be more consistent with the data, given that structural cues are also used in the resolution of agreement dependencies, for which our datasets show no participants with high cue weighting estimates.

The Relationship Between Latency Factor and Cue Weighting

We turn now to the relationship between reading speed and cue weighting. The possibility of a population-level correlation between latency factor and cue weighting is suggested by the empirical data of Dillon et al. (2013) and Jäger et al. (2020), as well as by the hypothesis that more skilled readers read more quickly and might apply syntactic constraints more strictly. We evaluated this question using all 13 datasets available. After fitting the cue-based retrieval model using ABC, a meta-analysis of the population-level correlations between the latency factor (a parameter indexing average reading speed in the model) and cue weighting showed a correlation of −0.26 (95% CrI [−0.38, −0.12]): The faster participants read, the higher they tend to weight the structural cue over the number cue. This preliminary result should be tested in a new, confirmatory experiment; however, it is suggestive first evidence for a relationship

between reading speed as an index of reading skill and how strongly participants adhere to grammatical constraints during reading.

Limitations of the Present Study, and Future Directions

Our method provides new ways of investigating individual differences in sentence comprehension; however, there are a number of limitations that should be addressed in follow-up studies.

First, we have focused on facilitatory interference effects in ungrammatical sentences in this article. The main reason behind this decision is that the evidence for facilitatory interference in ungrammatical sentences—that is, distractor-induced speedups—in the literature is relatively robust. By contrast, the evidence for inhibitory interference in grammatical sentences—that is, distractor-induced slowdowns—which is also predicted by the Lewis and Vasishth (2005) model, is much more mixed (Jäger et al., 2017). While it is important to understand how individual-level latency factors, cue weightings, and their correlation behave in grammatical sentences (e.g., *The bodybuilder who worked with the trainer[s] was complaining ...*), the LV05 model likely needs to be augmented to account for the full range of results. Work is underway to apply an augmented version of the LV05 model to both grammatical and ungrammatical configurations simultaneously to gain a more complete picture of the individual variability in agreement and reflexive constructions. Nevertheless, it is still worthwhile to investigate how well the base LV05 model captures participants' behavior across the different dependencies in ungrammatical sentences, as we have done in the present work.

An additional limitation comes from the fact that we used the individual-level estimates of the facilitatory interference effect from the published data as the basis for our simulations. These estimates come with a high degree of uncertainty. That is not an issue specific to the data we used: reading times are highly variable and studies are often underpowered (see appendix B of Jäger et al., 2017). The high uncertainty of the data increases the uncertainty of our simulation-based parameter estimates. One solution to this problem is to collect more data from each participant. This provides more accurate estimates of individual parameter values, even though longer experiments may result in adaptation to the linguistic manipulation and reduced differences between conditions (Fine et al., 2013). Finding the right balance here is a challenging task for future work.

Our implementation of the Lewis and Vasishth (2005) model assumes that individual-level parameter values are sampled from a unimodal Gaussian distribution that is centered around the population-level mean. This assumption constrains the allowed variability across individuals, which is assumed to be quantitative rather than qualitative in nature (Haaf & Rouder, 2019; Navarro et al., 2006). In future work, we plan to compare models under the following assumptions: (1) participant-level cue weighting comes from a unimodal Gaussian distribution, (2) participant-level cue weighting comes from a bimodal distribution such that a participant either has cue weighting 1 or cue weighting > 1, and (3) participant-level cue weighting is constant, that is, 1 for all participants. If models (1) and (2) are better than (3), then there are individual differences in cue weighting. If model (2) is better than model (1), then cue-weighting variation among participants is split into two groups: participants with cue weighting 1 and participants with higher cue weighting.

Regarding our choice of parameters to fit, there are alternative choices available, which may result in different conclusions. While the choice of cue weighting is based on claims in the literature, the choice of the latency factor as representing reading speed and, by extension, language experience and fluency, may be somewhat more contentious. The latency

factor is one of the most frequently estimated parameters in the ACT-R literature (Wong et al., 2010), but it is usually used to account for differences between studies rather than differences between individuals (Anderson et al., 1998). Other candidate sources of individual differences in ACT-R and the LV05 model are the activation noise parameter, the goal activation parameter, or the default action time parameter, which have been used in work on aphasia (Mätzig et al., 2018; Patil et al., 2016).

Additionally, reading speed as a global parameter, as indexed by the latency factor, is a multifaceted concept that can be measured with different tasks, only a subset of which may correlate with reading comprehension (Gerst et al., 2021; Jackson & McClelland, 1979). Ideally, in order to test the proposed connection between reading speed and cue weighting, participants' reading speed should be measured independently, in a separate task. Another possibility is to compute average reading speed for filler sentences (Traxler et al., 2012), which we plan to do in future work.

We have assumed a simple relationship between reading speed and language skill: higher speed should be associated with more thorough syntactic processing. However, the data on this relationship are not as straightforward as one might hope. For instance, Roberts and Felser (2011) found lower comprehension accuracy for the faster readers in their native-speaker group for some garden-path sentences. Kaan et al. (2015) observed faster average reading in English for Dutch L2 learners than for English native speakers, as well as weaker online sensitivity to agreement errors for fast readers, both for native speakers and L2 learners. Findings like these suggest that reading speed does not necessarily correlate positively with experience or with comprehension. Readers may trade in accuracy for speed during reading, in accordance with whether their goal is detailed, holistic comprehension or something else, such as skimming the text for a particular type of information (Rayner et al., 2016). Reading may also become faster when memory retrievals fail due to lower working memory capacity (Nicenboim et al., 2016) or when certain aspects of the structural and semantic representation of the sentence are strategically left underspecified (Swets et al., 2008; von der Malsburg & Vasishth, 2013). Even if there is a stable underlying relationship between language experience, reading speed and comprehension, it may thus be obscured by varying task demands and differences in intrinsic and experiment-specific motivation (Schiefele et al., 2012). This is a challenge for almost all implemented models of sentence comprehension: Only a few implemented models (Logačev & Vasishth, 2015, 2016) explicitly address task effects on processing. Future work should investigate ways of including task effects in the LV05 model.

It is also not clear whether each individual has characteristic, fixed values for cue weighting and latency factor *within* a single study, as we have assumed here. This question can be answered by evaluating test-retest reliability, that is, by running the same experiment twice with each participant and checking whether the latency factor and cue weighting estimates from the first and second studies correlate. The reliability of individual differences in reading measures has recently started receiving more attention, as it is often unclear whether the observed differences between participants represent stable individual characteristics. While global reading speed has relatively high reliability, low reliability has been reported for the participant-level effects of some linguistic manipulations (Cunnings & Fujita, 2020; James et al., 2018; Staub, 2021), casting doubt on the assumption that individual differences can reliably be estimated in standard psycholinguistic experiments. Addressing this issue in the context of computational modeling is a further challenge. An important achievement of the present work is that we make these challenges explicit by using a computationally implemented model. Such a computational approach makes hypotheses more constrained and falsifiable than with a merely verbal model.

CONCLUSION

Within sentence processing research, this work is, to our knowledge, the first attempt at simultaneously estimating multiple individual-level parameters and their correlation using approximate Bayesian computation. We have presented a novel investigation of the cue-weighting hypothesis and its implications for dependency completion in sentence comprehension. The theoretical insight from our modeling approach is that there is variation among individual speakers in the application of linguistic constraints, and that psycholinguistic hypotheses should be evaluated with regard to whether they hold for all, some, or none of the sampled participants. Furthermore, there is some indication that cue weighting may be tied to language experience, as reflected in average reading speed. While there are still many open questions, the computational and statistical approach pioneered by Turner and Van Zandt (2014) and others that we have applied here is broadly applicable across cognitive science; it can be easily adapted to different computational modeling settings.

ACKNOWLEDGMENTS

The modeling work would not have been possible without the raw data provided by Brian W. Dillon, Sol Lago, and Matt Wagers; our heartfelt thanks to them. We also thank two anonymous reviewers for insightful comments.

FUNDING INFORMATION

HY, Deutscher Akademischer Austauschdienst (https://dx.doi.org/10.13039/501100001655), Award ID: 91730718. SV, Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID: 317633480, SFB 1287.

AUTHOR CONTRIBUTIONS

HY: Conceptualization: Equal; Formal analysis: Lead; Methodology: Lead; Visualization: Lead; Writing - Original Draft: Equal; Writing - Review & Editing: Equal. DP: Conceptualization: Equal; Writing - Original Draft: Equal; Writing - Review & Editing: Equal. GS: Conceptualization: Equal; Formal analysis: Supporting; Methodology: Supporting; Supervision: Supporting; Writing - Original Draft: Equal; Writing - Review & Editing: Equal. BWD: Data Curation: Lead; Resources: Supporting; Writing - Review & Editing: Equal. SV: Conceptualization: Equal; Formal analysis: Supporting; Methodology: Supporting; Supervision: Lead; Writing - Original Draft: Equal; Writing - Review & Editing: Equal.

REFERENCES

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380. https://doi.org/10.1006/jmla.1997.2553

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press. https://doi.org/10.4324/9781315805696

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001, PubMed: 24403724

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 [Computer software manual] (R package version 1.0-6). R Foundation for Statistical Computing.

Chomsky, N. (1981). *Lectures on government and binding*. Foris.

Cunnings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, *20*(4), 659–678. https://doi.org/10.1017/S1366728916000675

Cunnings, I., & Fujita, H. (2020). Quantifying individual differences in native and nonnative sentence processing. *Applied Psycholinguistics*, 1–21. https://doi.org/10.1017/S0142716420000648

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139. https://doi.org/10.1016/j.jml.2014.05.006

Cunnings, I., & Sturt, P. (2018). Retrieval interference and sentence interpretation. *Journal of Memory and Language*, *102*, 16–27. https://doi.org/10.1016/j.jml.2018.05.001

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition,

language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277–299. https://doi.org/10.1207/s1532799xssr1003_5

Danker, J. F., Fincham, J. M., & Anderson, J. R. (2011). The neural correlates of competition during memory retrieval are modulated by attention to the cues. *Neuropsychologia*, *49*(9), 2427–2438. https://doi.org/10.1016/j.neuropsychologia.2011.04.020, PubMed: 21549721

Dillon, B. W., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. https://doi.org/10.1016/j.jml.2013.04.003

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. In S. Kepser & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives* (pp. 145–164). De Gruyter Mouton. https://doi.org/10.1515/9783110197549.145

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *43*(12), Article e12800. https://doi.org/10.1111/cogs.12800, PubMed: 31858626

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*(10), Article e77661. https://doi.org/10.1371/journal.pone.0077661, PubMed: 24204909

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 1–32.

Gerst, E. H., Cirino, P. T., Macdonald, K. T., Miciak, J., Yoshida, H., Woods, S. P., & Gibbs, M. C. (2021). The structure of processing speed in children and its impact on reading. *Journal of Cognition and Development*, *22*(1), 1–24. https://doi.org/10.1080/15248372.2020.1862121, PubMed: 33519305

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin and Review*, *26*(3), 772–789. https://doi.org/10.3758/s13423-018-1522-x, PubMed: 30251148

Jackson, M. D., & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology: General*, *108*(2), 151–181. https://doi.org/10.1037/0096-3445.108.2.151, PubMed: 528903

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339. https://doi.org/10.1016/j.jml.2017.01.004

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*, Article 104063. https://doi.org/10.1016/j.jml.2019.104063, PubMed: 33100507

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*, 155–181. https://doi.org/10.1016/j.jml.2018.05.006, PubMed: 30713367

Jenkins, J. R., Fuchs, L. S., Van Den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, *95*(4), 719–729. https://doi.org/10.1037/0022-0663.95.4.719

Joo, S. J., Tavabi, K., Caffarra, S., & Yeatman, J. D. (2021). Automaticity in the reading circuitry. *Brain and Language*, *214*, Article 104906. https://doi.org/10.1016/j.bandl.2020.104906, PubMed: 33516066

Kaan, E., Ballantyne, J. C., & Wijnen, F. (2015). Effects of reading speed on second-language sentence processing. *Applied Psycholinguistics*, *36*(4), 799–830. https://doi.org/10.1017/S0142716413000519

Kangasrääsiö, A., Jokinen, J. P., Oulasvirta, A., Howes, A., & Kaski, S. (2019). Parameter inference for computational cognitive models with approximate Bayesian computation. *Cognitive Science*, *43*(6), Article e12738. https://doi.org/10.1111/cogs.12738, PubMed: 31204797

Kreiner, H., Garrod, S., & Sturt, P. (2013). Number agreement in sentence comprehension: The relationship between grammatical and conceptual factors. *Language and Cognitive Processes*, *28*(6), 829–874. https://doi.org/10.1080/01690965.2012.667567

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42–73. https://doi.org/10.1016/j.jml.2011.03.002, PubMed: 21709808

Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing* (PhD thesis). University of Maryland, College Park, MD.

Kwon, N., & Sturt, P. (2016). Attraction effects in honorific agreement in Korean. *Frontiers in Psychology*, *7*, Article 1302. https://doi.org/10.3389/fpsyg.2016.01302, PubMed: 27630594

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149. https://doi.org/10.1016/j.jml.2015.02.002

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000_25, PubMed: 21702779

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., van het Nederend, M. L., Burchert, F., Stadie, N., Caplan, D., & Vasishth, S. (2021). A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science*, *45*(4), Article e12956. https://doi.org/10.1111/cogs.12956, PubMed: 33877698

Logačev, P., & Vasishth, S. (2015). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*(2), 266–298. https://doi.org/10.1111/cogs.12228, PubMed: 25823920

Logačev, P., & Vasishth, S. (2016). Understanding underspecification: A comparison of two computational implementations. *The Quarterly Journal of Experimental Psychology*, *69*(5), 996–1012. https://doi.org/10.1080/17470218.2015.1134602, PubMed: 26960441

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*(2), 123–146. https://doi.org/10.1080/1057356970130203

Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., & Burchert, F. (2018). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, *10*(1), 161–174. https://doi.org/10.1111/tops.12323, PubMed: 29356427

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123. https://doi.org/10.1023/A:1005184709695, PubMed: 10709178

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122. https://doi.org/10.1016/j.jmp.2005.11.006

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology*, *7*, 280. https://doi.org/10.3389/fpsyg.2016.00280, PubMed: 27014113

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34. https://doi.org/10.1016/j.jml.2017.08.004

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, *18*(1), 5–19. https://doi.org/10.1007/BF01069043, PubMed: 2647962

Paape, D., Avetisyan, S., Lago, S., & Vasishth, S. (2020). *Modeling misretrieval and feature substitution in agreement attraction: A computational evaluation*. PsyArXiv. https://doi.org/10.31234/osf.io/957e3

Paape, D., Avetisyan, S., Lago, S., & Vasishth, S. (2021). Modeling misretrieval and feature substitution in agreement attraction: A computational evaluation. *Cognitive Science*, *45*(8), Article e13019. https://doi.org/10.1111/cogs.13019, PubMed: 34379348

Palestro, J. J., Sederberg, P. B., Osth, A. F., Van Zandt, T., & Turner, B. M. (2018). *Likelihood-free methods for cognitive science*. Springer. https://doi.org/10.1007/978-3-319-72425-6

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290. https://doi.org/10.1016/j.jml.2017.01.002

Patil, U., Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2016). A computational evaluation of sentence processing deficits in aphasia. *Cognitive Science*, *40*(1), 5–50. https://doi.org/10.1111/cogs.12250, PubMed: 26016698

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*(5 Series II), 574–590. https://doi.org/10.1111/j.2164-0947.1962.tb01433.x, PubMed: 14489538

Rayner, K., Schotter, E. R., Masson, M. E., Potter, M. C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, *17*(1), 4–34. https://doi.org/10.1177/1529100615623267, PubMed: 26769745

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics*, *32*(2), 299–331. https://doi.org/10.1017/S0142716410000421

Rouder, J. N., & Haaf, J. M. (2021). Are there reliable qualitative individual difference in cognition? *Journal of Cognition*, *4*(1), Article 52. https://doi.org/10.5334/joc.131, PubMed: 34514317

Samuels, S. J., & Flor, R. F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*(2), 107–121. https://doi.org/10.1080/1057356970130202

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, Article 104038. https://doi.org/10.1016/j.jml.2019.104038

Schiefele, U., Schaffner, E., Möller, J., & Wigfield, A. (2012). Dimensions of reading motivation and their relation to reading behavior and competence. *Reading Research Quarterly*, *47*(4), 427–463.

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press. https://doi.org/10.1201/9781315117195

Sohn, M.-H., Anderson, J. R., Reder, L. M., & Goode, A. (2004). Differential fan effect and attentional focus. *Psychonomic Bulletin & Review*, *11*(4), 729–734. https://doi.org/10.3758/BF03196627, PubMed: 15581125

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*(11), 1497–1524. https://doi.org/10.1016/j.lingua.2004.07.002

Staub, A. (2021). How reliable are individual differences in eye movements in reading? *Journal of Memory and Language*, *116*, Article 104190. https://doi.org/10.1016/j.jml.2020.104190

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*(3), 542–562. https://doi.org/10.1016/S0749-596X(02)00536-3

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, *36*(1), 201–216. https://doi.org/10.3758/MC.36.1.201, PubMed: 18323075

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual differences in eye-movements during reading: Working memory and speed-of-processing effects. *Journal of Eye Movement Research*, *5*(1). https://doi.org/10.16910/jemr.5.1.5

Turner, B. M., & Van Zandt, T. (2014). Hierarchical approximate Bayesian computation. *Psychometrika*, *79*(2), 185–209. https://doi.org/10.1007/s11336-013-9381-x, PubMed: 24297436

Underwood, G., Hubbard, A., & Wilkinson, H. (1990). Eye fixations predict reading comprehension: The relationships between reading skill, reading speed, and visual inspection. *Language and Speech*, *33*(1), 69–81. https://doi.org/10.1177/002383099003300105, PubMed: 2283921

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. Journal of Experimental Psychology. *Learning, Memory, and Cognition*, *33*(2), 407–430. https://doi.org/10.1037/0278-7393.33.2.407, PubMed: 17352621

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263. https://doi.org/10.1016/j.jml.2011.05.002, PubMed: 21927535

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712. https://doi.org/10.1080/03640210802066865, PubMed: 21635350

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, *23*(11), 968–982. https://doi.org/10.1016/j.tics.2019.09.003, PubMed: 31668586

von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545–1578. https://doi.org/10.1080/01690965.2012.728232

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. https://doi.org/10.1016/j.jml.2009.04.002

Wong, T. J., Cokely, E. T., & Schooler, L. J. (2010). An online database of ACT-R parameters: Towards a transparent community-based approach to model development. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 282–286). Drexel University.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language, 108*(1), 40–55. https://doi.org/10.1016/j.bandl.2008.10.002, PubMed: 19007980

Zhang, Z., Jiang, K., Liu, H., & Oh, I.-S. (2017). Bayesian meta-analysis of correlation coefficients through power prior. *Communications in Statistics—Theory and Methods*, *46*(24), 11988–12007. https://doi.org/10.1080/03610926.2017.1288251

APPENDIX
MODEL PREDICTIONS CONDITIONAL ON PARAMETER ESTIMATES FROM DILLON ET AL.'S (2013) DATASET

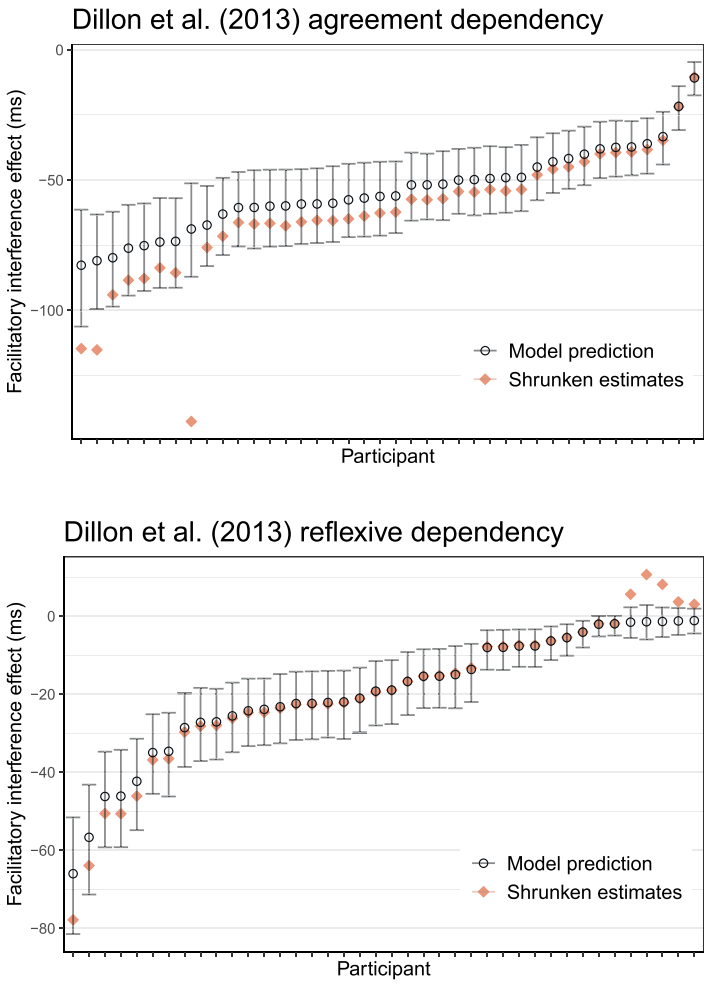


Figure A1. Posterior predicted values for individuals in agreement and reflexive dependencies, derived from the cue-based retrieval model after estimating individual-level parameters for cue weighting. Shown are the posterior mean and 95% credible interval, along with the estimate of the shrunken mean for each participant.

Chapter 5

Article III

Individuals differ cross-linguistically in cue weighting: A computational evaluation of cue-based retrieval in sentence processing

Himanshu Yadav, Garrett Smith, Daniela Mertzen, Ralf Engbert, and Shravan Vasishth

Proceedings of the Annual Meeting of the Cognitive Science Society, 44, 2022

Link: https://escholarship.org/uc/item/35m5q9rx

Individuals differ cross-linguistically in cue weighting: A computational evaluation of cue-based retrieval in sentence processing

Himanshu Yadav (hyadav@uni-potsdam.de)
University of Potsdam, 14476 Potsdam, Germany

Garrett Smith (gasmith@uni-potsdam.de)
University of Potsdam, 14476 Potsdam, Germany

Daniela Mertzen (daniela.mertzen@uni-potsdam.de)
University of Potsdam, 14476 Potsdam, Germany

Ralf Engbert (engbert@uni-potsdam.de)
University of Potsdam, 14476 Potsdam, Germany

Shravan Vasishth (vasishth@uni-potsdam.de)
University of Potsdam, 14476 Potsdam, Germany

Abstract

Cue-based retrieval theories of sentence processing assume that subject-verb dependencies are resolved through a content-addressable search in memory. The model assumes that multiple nouns with similar syntactic or semantic features increase dependency completion difficulty. English eyetracking data (reading) are consistent with model predictions; interestingly, a similar experiment with German–a language marking case overtly–suggests that only syntactic features affect dependency completion difficulty. Why would German show different behavior than English? Using a computational implementation of the cue-based retrieval model and model comparison using Bayes factors, we show that the reason is systematic variation at the individual-participant level: German participants overwhelmingly give higher weighting to syntactic cues over semantic cues, whereas English participants mostly give equal weighting to syntactic and semantic cues. The richer morphosyntax of German leads to syntactic cues being favoured; if such cues are largely absent (as in English) the parser relies on both cue types equally.

Keywords: Similarity-based interference; cue-based retrieval; individual differences

Introduction

Comprehending a sentence requires the reader to correctly figure out who did what to whom. This process of identifying the syntactic relations between words is called dependency completion. A well-established claim in sentence processing is that dependency completion between a verb and its associated subject is driven by a cue-based retrieval process (Lewis & Vasishth, 2005; McElree, 2000; Van Dyke, 2007). Under the cue-based retrieval account, the target noun is identified via a content-addressable search in memory based on feature specifications at the verb, such as [subject], called retrieval cues. When multiple nouns in memory match the retrieval cues, it is difficult to identify the target noun, which leads to a slowdown in retrieval times at the verb compared to a situation where only one noun matches the retrieval cues.

For example, in sentence (a) below, both the nouns *the resident* and *the neighbour* are in subject position, i.e., they both match the retrieval cue [subject], compared to sentence (b) where only one noun *the resident* matches the [subect] cue. The reading times at the verb *was complaining* are predicted to be slower in sentence (a) compared to sentence (b). This predicted effect is called *syntactic interference* (Van Dyke, 2007; Van Dyke & Lewis, 2003).

(a) . . . the resident who said that the neighbour was dangerous was complaining . . .

(b) . . . the resident who was living near the dangerous neighbor was complaining . . .

Similarly, when multiple nouns match the verb's semantic cues, such as [animate], they are assumed to cause *semantic interference* (Van Dyke, 2007). For example, in sentence (c) where both *the resident* and *the neighbour* are animate, the retrieval times at the verb are predicted to be slower than sentence (d), where only *the resident* is animate.

(c) . . . the resident who said that the neighbour was dangerous was complaining . . .

(d) . . . the resident who said that the warehouse was dangerous was complaining . . .

The predicted syntactic and semantic interference effects are consistently found in English reading studies (Van Dyke, 2007; Van Dyke & Lewis, 2003; Arnett & Wagers, 2017; Van Dyke & McElree, 2011). In a recent cross-linguistic study (Mertzen, Paape, Dillon, Engbert, & Vasishth, 2021), both syntactic and semantic interference were observed in English, but in German, only syntactic interference was observed at the verb (see Figure 1). Semantic interference was absent at the verb and appeared only later in the post-verbal region.

In sum, at the critical region (the verb phrase *was complaining*), syntactic interference predicted by the cue-based retrieval account is observed in both English and German, but semantic interference is observed only in English. The default assumption in cue-based retrieval models is that syntactic and semantic cues are used in the same way, so the magnitude of semantic and syntactic interference is predicted

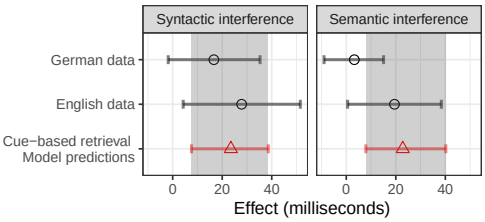


Figure 1: Syntactic and semantic interference predicted by a cue-based retrieval model (shaded areas; Lewis & Vasishth, 2005) compared with the observed effects in English and German from Mertzen et al. (2021); the effects are estimated from regression path durations at the verb. The error bars show 95% credible intervals of predicted or observed effects.

to be the same. The absence of semantic interference in German is, therefore, puzzling.

Mertzen et al. propose an explanation: a noun's syntactic cue is weighted higher than its semantic cue in German, while the two cues are weighted equally in English. The high weighting for syntactic cues in German could be due to the overt case marking on the nouns; it is possible that the overt case marking is highly reliable in identifying the grammatical functions of the nouns.

The cue weighting proposal is not entirely new. Several researchers have hypothesized that syntactic cues may be weighted more strongly over non-syntactic cues in processing antecedent-reflexive dependencies (Dillon, Mishler, Sloggett, & Phillips, 2013; Cunnings & Sturt, 2014; Kush, 2013; Parker & Phillips, 2017). A major limitation of these cue-weighting proposals is that they aim to explain the data averaged across all the participants. However, it is possible that individual differences in cue weighting exist. For example, a recent study on English (Yadav et al., 2021) showed that only one-third of the participants weigh syntactic cues more strongly over number cues in processing antecedent-reflexive dependencies. This result implies that the claim based on the average behavior holds only for a small subset of participants. This study demonstrates that the average behavior may mask theoretically important information which can only be revealed by modeling individual-level differences.

Modeling individual differences in syntactic and semantic interference in English and German might reveal a more nuanced picture of cue weighting differences among individuals and among the two language groups. For instance, it is possible that only a small subset of German participants, who have high weighting for the syntactic cue, is responsible for the absence of semantic interference in German. Therefore, the cue weighting hypothesis — that syntactic cues are weighted more strongly over semantic cues in German, but not in English — should be formulated for individual participants. The important questions to be asked are: (1) whether individuals differ in how they weight syntactic cues relative to semantic cues, and (2) whether individual German participants differ from individual English participants in cue weighting.

We test these questions by implementing two hierarchical models based on the Lewis and Vasishth (2005) cue-based retrieval model: (i) *the equal cue-weighting model*, which assumes that all the individuals have equal weights for syntactic and semantic cues, and (ii) *the varying cue-weighting model*, which assumes that individuals may differ in how highly they weight syntactic cues over semantic cues. The models are fitted to data from Mertzen et al. (2021) and then compared using Bayes factors (Rouder, Haaf, & Vandekerckhove, 2018).

The main finding is that there are cross-linguistic differences in individual-level cue-weighting: most German participants have higher weights for syntactic cues over semantic cues, while most English participants have equal weights for syntactic and semantic cues.

We first present the two individual difference models. Next, we quantify relative evidence for the two models and show the individual-level cue weighting estimates. We then discuss the broader implications of the work and conclude.

Two models of individual-level cue weighting

We implement two hierarchical models that differ in their assumption about the distribution of individual-level cue weighting. The models are implemented within the cue-based retrieval framework of Lewis and Vasishth (2005).

The Lewis and Vasishth (2005) model (see Engelmann, Jäger, & Vasishth, 2020, for the latest implementation) assumes that each noun phrase that matches a retrieval cue receives a certain amount of activation (see Figure 2). The total activation of a noun phrase i is given by

$$A_i = B_i + \sum_{j=1}^{n} W_j S_{ji} + \varepsilon_i \quad (1)$$

where B_i is the baseline activation of the noun i determined by its past retrievals, and ε_i is Gaussian noise added to activation of the noun i, such that $\varepsilon_i \sim Normal(0, \sigma)$. The term $\sum_{j=1}^{n} W_j S_{ji}$ represents that the noun phrase i receives activation from all matching cues j depending on the associative strength S_{ji} between the cue j and the noun i, and the cue's weight W_j (Engelmann et al., 2020). The cue's weight is determined by a parameter called *cue weighting*. Cue weighting encodes the ratio of weights of syntactic cues and non-syntactic cues. Following Yadav et al. (2021), we assume that the cue weighting can have a value between 1 and 4, such that the cue weighting of 1 means equal weights for syntactic and semantic cues and the cue weighting of 4 means four times higher weight for the syntactic cue over semantic cues.

The Lewis and Vasishth model further assumes that a noun phrase with the highest activation gets retrieved for dependency completion. The retrieval time at the verb is determined by the activation level of the retrieved noun, A_i.

$$T = Fe^{-A_i} \quad (2)$$

where the latency factor F reflects overall reading speed and may, inter alia, include lexical access time, motor response

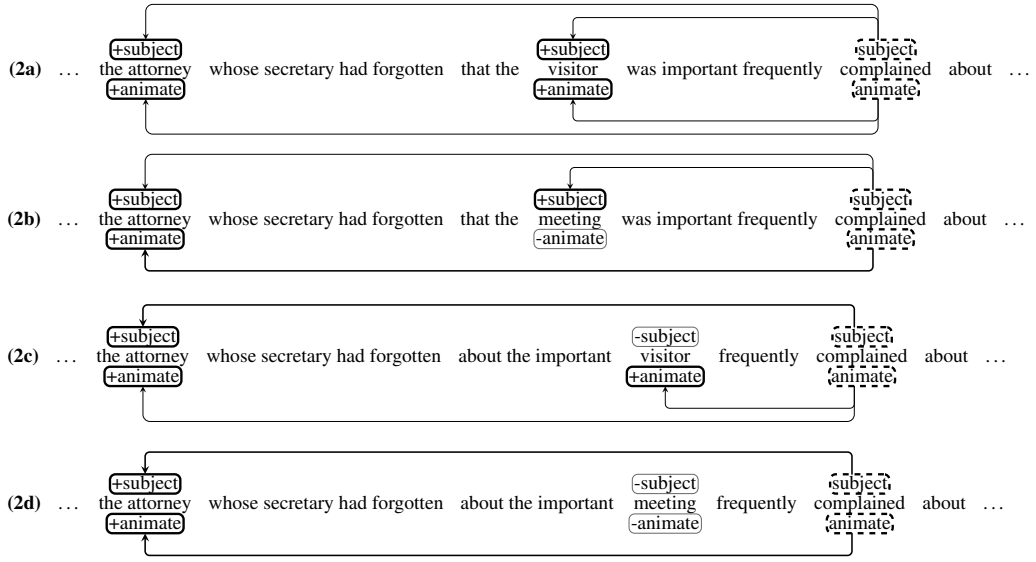


Figure 2: Activation received by the nouns based on cue-feature match, with a thick arrow denoting more activation spreading compared to a thin arrow. A dashed box represents a retrieval cue, a thick box represents a feature that matches a retrieval cue, and a thin box represents a feature that does not match a retrieval cue.

time, etc. The latency factor is commonly considered a free parameter in the model.

Figure 2 shows how activation spreads to each noun phrase in the four example conditions from Mertzen et al. (2021). In conditions (2a) and (2b), multiple nouns match the [subject] cue at the verb; as a result, the activation spread via the [subject] cue gets divided among these nouns. This is called the *fan effect* (Anderson et al., 2004; Schneider & Anderson, 2012). Due to the fan effect, the retrieval times at the verb in conditions (2a) and (2b) are predicted to be slower compared to conditions (2c) and (2d); this slowdown is referred to as syntactic interference. Similarly, due to the fan effect of the [+animate] feature in conditions (2a) and (2c), the retrieval times in (2a) and (2c) are predicted to be slower than in (2b) and (2d), which is called semantic interference.

Based on the equations 1 and 2, the model predicts syntactic and semantic interference, (X_{syn}, X_{sem}) as a function of cue weighting W and latency factor F,

$$(X_{syn}, X_{sem}) \sim Model(W, F) \tag{3}$$

The magnitude of both syntactic and semantic interference increases linearly with an increase in latency factor. But the cue weighting affects only semantic interference: the magnitude of semantic interference decreases with an increase in cue weighting. This is because with the increase in cue weighting, the semantic cue gets weaker, and consequently, the fan effect caused by the semantic cue gets weaker, which leads to the decrease in semantic interference.

We implement two hierarchical models that predict syntactic and semantic interference for each individual participant as a function of individual-level cue weighting and latency factor. The models make different assumptions about how the cue weighting varies among individuals, which we discuss next.

The equal cue-weighting model

The equal cue-weighting model assumes that all participants have equal weighting for syntactic and semantic cues.

Suppose that $(X_{syn_{j,g}}, X_{sem_{j,g}})$ represent syntactic and semantic interference effects for a participant j from language g.

$$(X_{syn_{j,g}}, X_{sem_{j,g}}) \sim Model(W_{j,g}, F_{j,g}) \tag{4}$$

where $W_{j,g}$ is the cue weighting and $F_{j,g}$ is the latency factor of the participant j of language g.

Under the equal cue-weighting model, all the participants regardless of their language have cue weighting equal to 1, i.e., they have equal weights for syntactic and semantic cues.

$$W_{j,g} = 1 \tag{5}$$

The individual-level latency factor $F_{j,g}$ is assumed to come from a normal distribution with population-level mean latency factor μ_{F_g} and population-level variance $\tau^2_{F_g}$ for language g:

$$F_{j,g} \sim Normal_{lb=0.05}(\mu_{F_g}, \tau^2_{F_g}) \tag{6}$$

where $lb = 0.05$ represent a lower bound of 0.05 on latency factor values. We choose this lower bound because a latency factor of less than 0.05 would generate unreasonably fast reading times for an individual (see Jäger, Engelmann, & Vasishth, 2017, for a meta-analysis of reading times).

The varying cue-weighting model

The varying cue-weighting model assumes that participants may differ in weighting of syntactic cues over semantic cues.

Suppose that $(X_{syn_{j,g}}, X_{sem_{j,g}})$ represents the syntactic and semantic interference effects for a participant j from the language g.

$$(X_{syn_{j,g}}, X_{sem_{j,g}}) \sim Model(W_{j,g}, F_{j,g}) \tag{7}$$

The individual-level latency factor $F_{j,g}$ is assumed to come from the same distribution as shown in Equation 6.

Under the varying cue-weighting model, the cue weighting for the participant j of language g, i.e., $W_{j,g}$ comes from a normal distribution with population-level mean cue weighting μ_{W_g} and between-participant variance $\tau^2_{W_g}$:

$$W_{j,g} \sim Normal_{lb=1,ub=4}(\mu_{W_g}, \tau^2_{W_g}) \tag{8}$$

where $lb = 1, ub = 4$ constrains the individual-level cue weighting to be between 1 and 4. A cue weighting of 1 means equal weights for syntactic and semantic cues and a cue weighting of 4 means 4 times higher weight for the syntactic cue.

The population-level cue weighting parameters, the mean cue weighting μ_{W_g} and between-participant variance $\tau^2_{W_g}$, are the main parameters that make the varying cue-weighting model different from the equal weighting model. A comparative evaluation of the two models can be sensitive to the priors on these population-level cue weighting parameters. Following the recommendation in Schad et al. (2021), we choose a range of priors on mean cue weighting and between-participant variance in cue weighting so that we can compare the models under different assumptions about the distribution of cue weighting in the populations.

For the population-level mean cue weighting μ_{W_g}, we specify the following prior:

$$\mu_{W_g} \sim Normal_{lb=1,ub=4}(1, \sigma_m) \tag{9}$$

where $\sigma_m \in \{0.05, 0.1, 0.5, 1\}$. The different values of σ_m express our assumptions about possible values of mean cue weighting. For example, $Normal_{lb=1,ub=4}(1, 0.05)$ represents that the mean cue weighting is retricted to be very close to 1, while $Normal_{lb=1,ub=4}(1, 1)$ represents that the mean cue weighting is allowed to be somewhere between 1 and 3.

For the between-participant variance in cue weighting $\tau^2_{W_g}$, we use an inverse-gamma prior.

$$\tau^2_{W_g} \sim InvGamma(1, scale) \tag{10}$$

where $scale \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$. The different values of *scale* express our assumptions about how much variation in cue weighting is allowed across individuals.

We fit these two models of individual-level cue weighting on data from Mertzen et al. (2021) and compute their marginal likelihoods given the data.

Model comparison

Mertzen et al. investigated both semantic and syntactic interference in a single design across two languages, English and German. From their dataset, we obtain shrunken estimates of individual-level syntactic and semantic interference for each participant as shown in Figure 3.[1]

We fit the equal cue weighting model and the varying cue-weighting model on the individual-level interference effects using hierarchical Approximate Bayesian Computation (Turner & Van Zandt, 2014; Sisson, Fan, & Beaumont, 2018) and obtained the marginal likelihoods for the each model given the data.

We then quantified the evidence for the varying cue-weighting model against the equal cue-weighting model using the Bayes factors (Rouder et al., 2018; Schönbrodt & Wagenmakers, 2018). The Bayes factor in favor of a model $\mathcal{M}_1$ compared to a model $\mathcal{M}_2$, i.e., BF_{12} is computed as the ratio of the marginal likelihoods of $\mathcal{M}_1$ and $\mathcal{M}_2$. The Bayes factor BF_{12} represents the extent to which the model $\mathcal{M}_1$ is more likely than $\mathcal{M}_2$ given the data. Following the convention from Jeffreys (1939/1998), a Bayes factor value of larger than 10 is interpreted as strong evidence in favor of $\mathcal{M}_1$ and a value between 3 and 10 is interpreted as moderate evidence in favor of $\mathcal{M}_1$.

Figure 4 shows the estimated Bayes factor under each prior assumption about the population-level cue weighting. We find that the Bayes factors are larger than 3 when the mean cue weighing is assumed to be very close to 1, suggesting moderate evidence in favor of the varying cue-weighting model. Under the assumption that the mean cue weighting could lie in the range of 1 to 2 or 1 to 3, the Bayes factors are larger than 10, indicating strong evidence in favor of the varying cue-weighting model.

Overall, the Bayes factors suggest moderate to strong evidence for the varying cue-weighting model compared to the equal cue-weighting model.

Individual-level cue weighting estimates

The model comparison shows evidence in favor of the assumption that individuals differ in cue weighting. But how do they differ? What is the distribution of individual-level cue weighting in English and German? We can answer this using individual-level cue weighting estimates from the varying cue-weighting model.

Figure 5 shows the estimated posterior distribution of cue-weighting for each individual participant from English and German. We find that 85% of the German participants have cue weighting larger than 2 meaning that 85% of the German participants give at least two times higher weights to syntactic cues over semantic cues. And, 84% of English participants have cue weighting of less than 1.5, which means that 84%

[1] To obtain individual-level interference effects, we fit a Bayesian hierarchical model with varying intercepts and slopes for participants and items, where regression path durations are the dependent variable and conditions (syntactic vs semantic, feature-match vs mismatch) are sum-coded predictors (Schad et al., 2020).

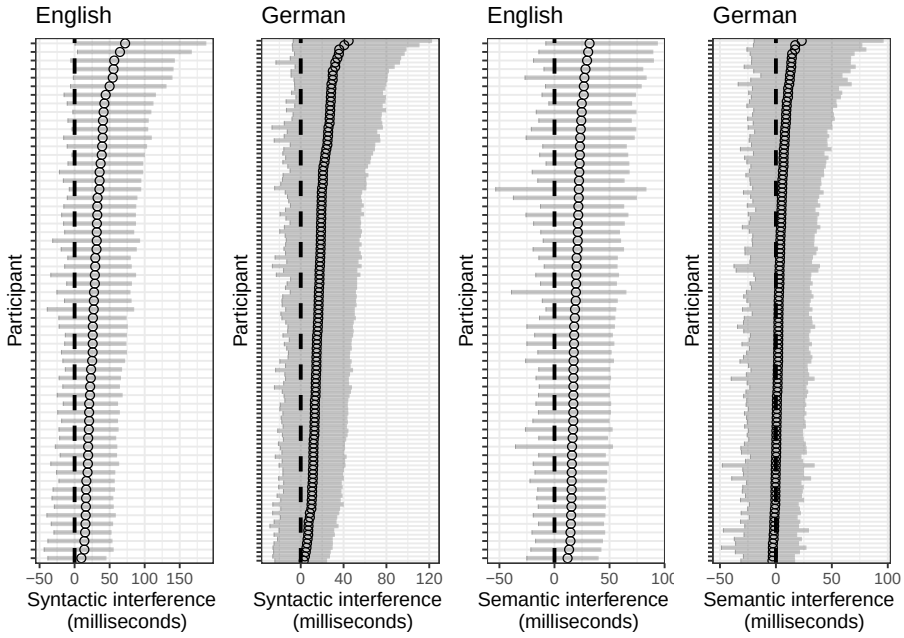


Figure 3: Individual-level syntactic and semantic interference effects from the Mertzen et al. (2021) data. Shown are the shrunken estimates from a Bayesian hierarchical model fit to regression path durations at the verb. English had 61 participants, German had 121 participants. The circles represent mean effects, the error bars represent 95% credible intervals of the effects.

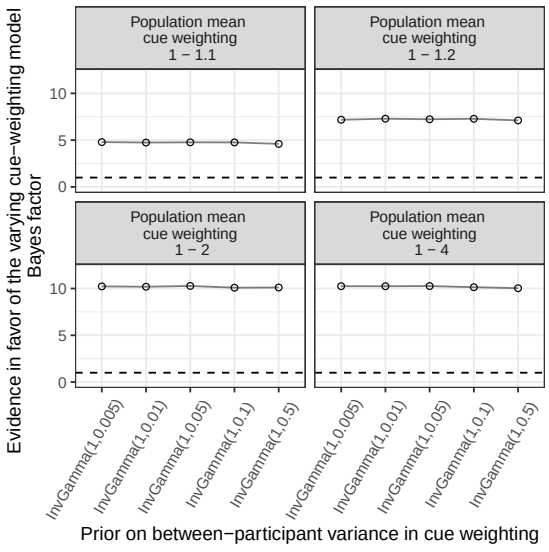


Figure 4: Estimated Bayes factors given different priors on the population-level mean cue weighting, μ_{W_g} and between-participant variance in cue weighting, $\tau^2_{W_g}$ (see Equation 8).

of English participants give approximately equal weights to syntactic and semantic cues.

In sum, the cue weighing estimates indicate that the most of the German participants have high weighting (> 2) for the syntactic cue, while the most of the English participants have equal weighting (≈ 1) for syntactic and semantic cues.

Discussion

Are syntactic and semantic retrieval cues weighted differently by English and German speakers? To answer this question, we implemented two hierarchical models, the equal cue-weighting model and the varying cue-weighting model. The equal cue-weighting model assumed that all English and German participants have equal weights for the syntactic and the semantic cues when retrieving a verb's subject from memory; the varying cue-weighting model assumed that individual participants can differ in how strongly they weight syntactic cues over semantic cues.

The models were evaluated on individual-level syntactic and semantic interference data from Mertzen et al. (2021). The model comparison and the model fits show that

1. There is moderate to strong evidence in favor of the varying cue-weighting model, suggesting that individuals vary in how they weight retrieval cues.

2. Most German participants give higher weights to syntactic cues over semantic cues, while most English participants give equal weights to syntactic and semantic cues.

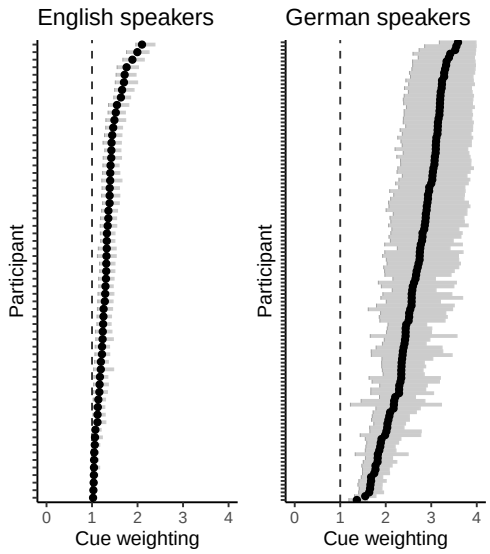


Figure 5: Individual-level cue weighting estimates from the varying cue-weigting model fitted to Mertzen et al. (2021) data. The solid circles represent mean cue weighting, the error bars represent 95% credible intervals of estimated values.

The results indicate that German speakers differ from English speakers in how they weight syntactic cues relative to non-syntactic cues. The conclusion is important for theories of sentence processing, because there is some independent support for the idea that the native speakers of a particular language may learn to use certain cues more strongly and reliably over the others (Dittmar et al., 2008; Sokolov, 1988; Bates et al., 1984). It is possible that German speakers weight the syntactic cues higher because the overt case marking in German is highly reliable in identifying the grammatical functions of the nouns in a sentence. For example, in the experimental items used in Mertzen et al. (2021), the grammatical role of every pre-verbal noun was identifiable either by (i) unambiguous case marking of the noun (if it was a masculine noun), or by (ii) the properties of its case-assigning head (verb or preposition).

A principled test of our conclusion would be in verifying whether the distribution of individual-level cue weighting in German and English is replicated in future experiments. If the inferred distribution — that most German speakers have high weighting for syntactic cues — holds for the language population, one would expect to see the same distribution of cue weighting in repeated experiments with larger samples of German participants. We plan to run a relatively large-sample-size study to test this prediction.

An interesting question that remains to be investigated is whether cue weighting is correlated with the general reading speed of an individual. There are reasons to believe that fast readers may weigh syntactic cues more strongly than slow readers (see Yadav et al., 2021). The strong weighting of syntactic cues in German speakers compared to English speakers might be associated with differences in their reading speed. Systematic experimental and modeling work is required to investigate the relationship between individual-level reading speed and cue weighting.

We have implemented only two models of individual differences in cue weighting, but one can explore other assumptions about how individuals vary in cue weighting. For example, it could be assumed that all German participants have a fixed cue weighting, which is different from English participants. Another assumption could be that only German participants vary in cue weighting while all English participants have equal cue weighting. It would be interesting to compare models under different assumptions about the distribution of individual-level cue weighting in German and English language populations.

A weakness of the current modeling work is that we do not have an independent measure of cue weighting for each individual; we can only infer it indirectly through reading times. The cue-weighting differences that are used to explain the observed individual differences in the data are estimated from the same data. It is possible that the individual-level cue weighting is overfitted to these data and that we may not get stable estimates of cue weighting for an individual in repeated experiments. A better approach would be to measure cue weighting independently for each participant on a separate processing task and then test the phenomenon of interest on the same group of participants. Using this approach, we can directly investigate whether the model can predict an individual's behavior based on their cue weighting. We plan to take this up in future work.

The current work reveals new insights about the constraints on processing subject-verb dependencies: The dependency between a verb and its associated subject is resolved via a cue-based retrieval process where the cues can be weighted differently by individuals depending on their native language. To our knowledge, this is the first investigation of cross-linguistic cue-weighting differences in a computational model of sentence comprehension. Our work contributes to understanding how different sources of linguistic information are employed during processing.

Acknowledgments

We thank the three anonymous reviewers for helpful suggestions. HY received funding from the Deutscher Akademischer Austauschdienst - DAAD, Programm ID: 57440921. SV received funding from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 317633480, SFB 1287.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–60.

Arnett, N., & Wagers, M. (2017). Subject encodings and retrieval interference. *Journal of Memory and Language*, *93*, 22–54.

Bates, E., MacWhinney, B., Caselli, C., Devescovi, A., Natale, F., & Venza, V. (1984). A cross-linguistic study of the development of sentence interpretation strategies. *Child development*, 341–354.

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139.

Dillon, B. W., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*, 85–103.

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child development*, *79*(4), 1152–1167.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2020). The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science*. Retrieved from 10.1111/cogs.12800

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316-339.

Jeffreys, H. (1939/1998). *The theory of probability*. Oxford University Press.

Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing*. Phd thesis, University of Maryland, College Park, MD.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.

Mertzen, D., Paape, D., Dillon, B., Engbert, R., & Vasishth, S. (2021). Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data.

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part iv: Parameter estimation and bayes factors. *Psychonomic bulletin & review*, *25*(1), 102–113.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). *Workflow techniques for the robust use of bayes factors.* (draft)

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038.

Schneider, D. W., & Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology*, *64*(3), 127–160.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review*, *25*(1), 128–142.

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate bayesian computation*. CRC Press.

Sokolov, J. L. (1988). Cue validity in hebrew sentence comprehension. *Journal of Child Language*, *15*(1), 129–155.

Turner, B. M., & Van Zandt, T. (2014). Hierarchical approximate bayesian computation. *Psychometrika*, *79*(2), 185–209.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(2), 407–430.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.

Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2021). Individual differences in cue weighting in sentence comprehension: An evaluation using Approximate Bayesian Computation. *Open Mind*. (Provisionally accepted)

Chapter 6

Discussion

In the articles presented in the previous chapters, I have addressed two questions. First, which theory can best explain the observed range of number agreement effects in both grammatical and ungrammatical sentences? Second, why does the antecedent-reflexive dependency not show an agreement attraction effect?

My modeling work has revealed three key insights: (a) A hybrid model that assumes both cue-based retrieval and probabilistic feature distortion can best explain the number agreement effects observed for subject-verb number agreement dependencies, (b) Some participants weigh syntactic cue higher than the number cue in processing reflexive dependencies which leads to the absence of the number attraction effect in these dependencies, and (c) Individual difference modeling is important for theory development.

In this chapter, I discuss the above three insights in detail. In section 6.1, I propose a general theory of dependency completion processes based on the insights (a) and (b). In section 6.2, I discuss the theoretical significance of modeling individual differences in sentence comprehension.

6.1 The distortion-retrieval theory of sentence comprehension

The cue-based retrieval (Lewis and Vasishth, 2005, Lewis et al., 2006, McElree, 2003) is a well-established theory of how the comprehender identifies and links the linguistically related words in a sentence to interpret the intended message. The key assumption is that the dependencies between the words, such as dependencies between verbs and their associated subjects, are resolved via a content-addressable search in memory. The theory has been invoked to explain a broad range of empirical data from a variety of constructions, including subject-verb non-agreement dependencies (Mertzen et al., 2022, Van Dyke, 2007, Van Dyke and McElree, 2011), number agreement dependencies Dillon et al. (2013), Jäger et al. (2020), Lago et al. (2015), Wagers et al. (2009), plausibility mismatch configurations (Cunnings and Sturt, 2018), and negative polarity item licensing (Drenhaus et al., 2005, Vasishth et al., 2008).

However, as we discussed in our articles, there are at least two empirical challenges to the cue-based retrieval account: (i) the grammatical sentences' data from subject-verb number agreement dependencies, and (ii) the data from antecedent-reflexive dependencies. The theory in its original form fails to explain these data. But as we demonstrated, a few modifications in the assumptions of a cue-based retrieval model could account for these two datasets without losing the original empirical coverage of the model. A cue-based retrieval model that allows probabilistic feature distortion and the differential cue weighting assumption can explain the observed range

of effects for subject-verb number agreement dependencies and antecedent-reflexive dependencies. This is the main cumulative result of the modeling work presented in my three articles.

The result brings a fresh perspective to our understanding of cognitive processes that underlie sentence comprehension in humans. That is, a theory of sentence comprehension needs to incorporate the following three assumptions: (i) the representations stored in memory undergo probabilistic feature distortion, (ii) dependency completion is driven by a content-addressable search in memory, and (iii) the linguistic cues used for dependency completion can be weighted differentially.

I first discuss these three assumptions in detail followed by their implementation in the cue-based retrieval model of Lewis and Vasishth (2005). The model allows us to unify these assumptions within its architecture using only three free parameters. I call this modified model *the distortion-retrieval model*. After presenting the formal description of the model, I discuss the model predictions for different construction types studied in sentence comprehension.

6.1.1 Three important assumptions for a theory of sentence comprehension

The combined results from my three articles indicate that a theory of sentence comprehension needs to incorporate the assumptions of probabilistic representation distortion, cue-based retrieval, and differential cue-weighting. These assumptions find independent empirical support in psycholinguistics, and more generally in working memory research. I discuss the empirical and theoretical background of the three assumptions below.

(a) Probabilistic distortion of representations stored in memory

Comprehending a sentence requires the reader to temporarily maintain the relevant linguistic chunks in memory. The original cue-based retrieval theory assumes that the feature representation of the linguistic units stored in memory remains intact, i.e., the representations do not change over time.[1] The results from Article I provide a strong basis to consider the *probabilistic feature distortion assumption*: Some of the features of the nouns stored in memory can get changed or lost with time. For example, a singular noun can probabilistically change to plural when it is stored in memory along with a plural-marked noun. Two existing theories of sentence processing, the feature percolation theory (Bock and Eberhard, 1993, Eberhard, 1997) and the lossy-context surprisal theory (Futrell et al., 2020), already have the feature distortion assumption; however, they differ in terms of mechanism behind such a distortion. The feature percolation theory assumes that the number feature from a non-subject noun percolates to the subject noun probabilistically, and the noisy channel theory assumes that the features can get inserted or deleted over time constrained by information-theoretic principles. The exact nature of feature distortion remains an empirical question but the assumption has some independent support in memory literature. For example, when participants are asked to report the feature(s) of the target item that was recently presented along with a distractor item, then in a proportion of trials, participants make *swap errors*: they mistakenly report the features of the distractor item when probed about the target item (Bays, 2016, Bays et al., 2009, Scotti et al., 2021). Swap errors support the idea that features can migrate from one memory item to others. As we also show in Article I, a feature distortion assumption is necessary to explain the data from subject-verb number agreement dependencies. Thus, the probabilistic feature distortion is a well-motivated assumption and it would increase the empirical coverage of a sentence processing model.

[1] Although, the cue-based retrieval model of Lewis and Vasishth (2005) assumes that the accessibility of the representations can degrade over time, but it still assumes that the representations remain veridical in their content.

(b) Content-addressable search in memory

Sentence comprehension requires that the comprehender should figure out who did what to whom: the dependencies between linguistically related words must be completed. The cue-based retrieval theory assumes that the dependency completion is driven by a content-addressable search in memory. For instance, to complete a subject-verb dependency, a search is triggered at the verb based on its feature specification, such as [subject, plural], and a best-matching noun is retrieved from memory. The content-addressable search has been a key assumption in many theories that deal with working memory operations, including theories of sentence comprehension (Lewis and Vasishth, 2005, McElree, 2000), memory recall (Gillund and Shiffrin, 1984, Raaijmakers and Shiffrin, 1981), item recognition (Ratcliff, 1978), frequency judgments (Hintzman, 1984), etc. The assumption is supported by a range of empirical phenomena observed in sentence processing, such as agreement attraction (Avetisyan et al., 2020, Dillon et al., 2013, Lago et al., 2015, Tucker et al., 2015, Wagers et al., 2009), similarity-based interference (Mertzen et al., 2022, Van Dyke, 2007, Van Dyke and McElree, 2006, 2011), and semantic attraction (Cunnings and Sturt, 2018, Laurinavichyute and von der Malsburg, 2022). Our results in Article I also show that a content-addressable mechanism is important for explaining the data on subject-verb number agreement processing. Specifically, we find that a feature distortion assumption must be combined with the cue-based retrieval assumption to achieve the best fit. Given this overwhelming empirical support, the content-addressable search remains a very important assumption for modeling sentence comprehension.

(c) Differential weighting of the linguistic cues used for dependency completion

A widely-held assumption in sentence processing is that the comprehender combines different sources of linguistic information to resolve a dependency (Lewis and Vasishth, 2005, MacDonald et al., 1994, McElree, 2000, McRae et al., 1998). For instance, early constraint-based models (MacDonald et al., 1994, McRae et al., 1998) maintained that different types of constraints, including thematic fit, tense/voice information, frequency biases, etc., are simultaneously active in resolving local ambiguities in sentences like *the cop arrested by the detective was guilty of taking bribes*. The cue-based retrieval models Lewis and Vasishth (2005), McElree (2000) make a similar assumption: in order to complete a dependency, a set of linguistic cues such as case, number, and animacy, is used to search the co-dependents in memory. These linguistic cues are often assumed to be weighted equally, i.e., all types of cues are assumed to make an equal contribution in the retrieval process (but see Cunnings and Sturt, 2014, Dillon et al., 2013, Kush, 2013, Parker and Phillips, 2017). Our results from Article II and III suggest that syntactic cues might be weighted more strongly over the morphological and semantic cues in processing certain dependencies and in certain languages. Moreover, the individuals might differ in how they weigh different linguistic cues. These results call for considering the **differential cue weighting assumption**: the comprehender can weigh a particular linguistic cue more strongly over the others during dependency completion.

There is also independent support for the cue-weighting assumption that comes from the thematic role assignment studies in children and adults. In these studies, the participant is asked to identify the agent and the patient of the action in transitive sentences like *the cat/ball ate the ball/cat*. It is assumed the participants use different linguistic cues such as word order, case, and animacy to identify the correct thematic role of a noun. The cross-linguistic differences are observed in *cue-reliability*: English speakers strongly rely on word order, Italian speakers rely on agreement or semantic cues over word order cues, and German speakers rely on case marking cues (Bates et al., 1982, Dittmar et al., 2008, MacWhinney et al., 1984). Moreover, this differential cue reliability develops in children over time. For example, five-years-old German children use the

word order cues over case marking, but the seven-year-olds rely on case markers in identifying the agent-patient roles Dittmar et al. (2008).

Not only in sentence processing, cue weighting has also had some support in broader memory research. In a recall task, Sohn et al. (2004) showed that when the participants were trained on different types of information in the training session, they weighed retrieval cues differently in the main recall task. Similarly, Danker et al. (2011) observed that some individuals learn to use certain cues more effectively and reliably compared to other cues. Given this considerable empirical support for the cue-weighting assumption and its strong theoretical roots, the assumption should be considered for a theory of sentence processing.

6.1.2 The distortion-retrieval model: Implementation of the three key assumptions in a cue-based retrieval model

The three assumptions discussed above can be implemented in the cue-based retrieval model of Lewis and Vasishth (2005). The model can integrate these additional components without losing its original empirical coverage of sentence processing data. I would refer to this modified cue-based retrieval model as **the distortion-retrieval model**.

To summarize, the distortion-retrieval model makes the following key assumptions:

(1) **The cue-based retrieval assumption**: The dependency completion between a pair of words is driven by a content-addressable search in memory.

(2) **Feature distortion assumption**: The co-dependents stored in memory can undergo a probabilistic change in feature representation over time; Consequently, the cue-based retrieval process operates on the probabilistically distorted representation of chunks.

(3) **Cue-weighting assumption**: The cues used for searching a target chunk in memory may have different weighting. In processing sentences like *the bodybuilder who worked with the trainers injured themselves*, the c-command cue at *themselves* can have higher weight than the plural number cue. The cues with higher weights exert greater influence in identifying and completing the dependencies.

To include the above three assumptions, the unified distortion-retrieval model would have three free parameters: (a) the scaling parameter F, which represents the average reading speed for an individual or a population, (b) the distortion rate parameter θ, which determines the degree of feature distortion in the nouns when they are stored in memory, and (c) the cue weighting parameter W, the ratio of weights of two linguistic cues used in dependency completion. The formal description of the distortion-retrieval model is presented below.

The distortion-retrieval model assumes that the representation of a noun phrase i stored in memory can get distorted with probability θ. Suppose R_i is the veridical representation of a noun i stored in memory.[2] The probabilistically-distorted representation of the noun i in the k^{th} trial as a function of **distortion rate parameter** θ is given by

$$R_{i,k} = \begin{cases} R_{i,d} & \text{if } z_k = 1 \\ R_i & \text{if } z_k = 0 \end{cases} \text{ where } z_k \sim Bernoulli(\theta) \tag{6.1}$$

$R_{i,k}$ is the representation of the noun i in the k^{th} trial, $R_{i,d}$ is the non-veridical representation of the noun produced by the feature distortion process. For the feature distortion process, I assume

[2]The nouns stored in memory refer to the noun phrases that appear in a sentence before a retrieval site, e.g., a verb.

that certain features of a noun stored in memory can migrate to another noun in memory with a probability θ. What kind of features can migrate from one noun to another? In the current model, the probabilistic distortion is allowed for only those features that are deemed to be encoded in real-time in memory, e.g., number, gender, and case. A reasonable approximation of such features comes from the inflectional morphemes in a language: the features that are encoded using inflectional morphemes in a language can undergo probabilistic distortion.

Following the Lewis and Vasishth (2005) model (see Engelmann et al., 2019, for the latest implementation), the distortion-retrieval model assumes that each noun phrase that matches a retrieval cue receives a certain amount of activation (see Figure 6.1). The total activation of a noun phrase i in the trial k is given by

$$A_{i,k} = B_i + \sum_{j=1}^{n} W_j S_{ji,k} + \epsilon_k \tag{6.2}$$

where B_i is the baseline activation of the noun i determined by its past history of retrievals, and ϵ_k is Gaussian noise added to activation of the noun i in the k^{th} trial, such that $\epsilon_k \sim Normal(0, \sigma)$.

The term $\sum_{j=1}^{n} W_j S_{ji,k}$ in Equation eq:activation represents that the noun phrase i receives activation from all matching cues j depending on the associative strength $S_{ji,k}$ between the cue j and the noun i, and the cue's weight W_j Engelmann et al. (2019). The cue's weight W_j is determined by a parameter called *cue weighting*. Cue weighting encodes the ratio of weights of syntactic cues and non-syntactic cues. Similar to Article II, I assume that the cue weighting can have a value between 1 and 4, such that the cue weighting of 1 means equal weights for syntactic and non-syntactic cues and the cue weighting of 4 means four times higher weight for the syntactic cue over the other cue.

The associative strength $S_{ji,k}$ between a noun and a cue is determined by the number of nouns in memory that matches the cue j: As the number of nouns that matches the cue j increases, the associative strength between a noun i and cue j decreases. This is called the *fan effect* Anderson et al. (2004), Schneider and Anderson (2012). Since the cue-feature match in a trial depends on the feature representation of the nouns in a trial, the associative strength $S_{ji,k}$ would be the function of the noun's representation in the k^{th} trial, $R_{i,k}$ (see Equation 6.1).

The model further assumes that a noun with the highest activation gets retrieved for dependency completion. Suppose $A_{1,k}, A_{2,k}, ..., A_{m,k}$ represent the respective activation levels of the m nouns in memory, the activation of the noun that is retrieved from the memory in the k^{th} trial is given by

$$A_{r,k} = max_{i=1}^{m} A_{i,k} \tag{6.3}$$

The retrieval time in the k^{th} trial is determined by the activation level of the retrieved noun $A_{r,k}$.

$$T_k = Fe^{-A_{r,k}} \tag{6.4}$$

where the **latency factor** F reflects the overall reading speed and may, among others, include visual processing time, lexical access time, motor response time, etc. The latency factor is commonly considered a free parameter in the original cue-based retrieval model of Lewis and Vasishth (2005).

Figure 6.1 demonstrate the model's assumption of how activation spreads to each noun phrase in an example sentence when the retrieval process is triggered at the reflexive *themselves*. At the retrieval site *themselves*, a search is triggered in memory for the target antecedent noun: a noun that c-commands the reflexive *themselves* and has plural features. In a proportion of trials, the plural feature of the noun *trainers* can migrate to the noun *bodybuilder* causing the noun *bodybuilder*

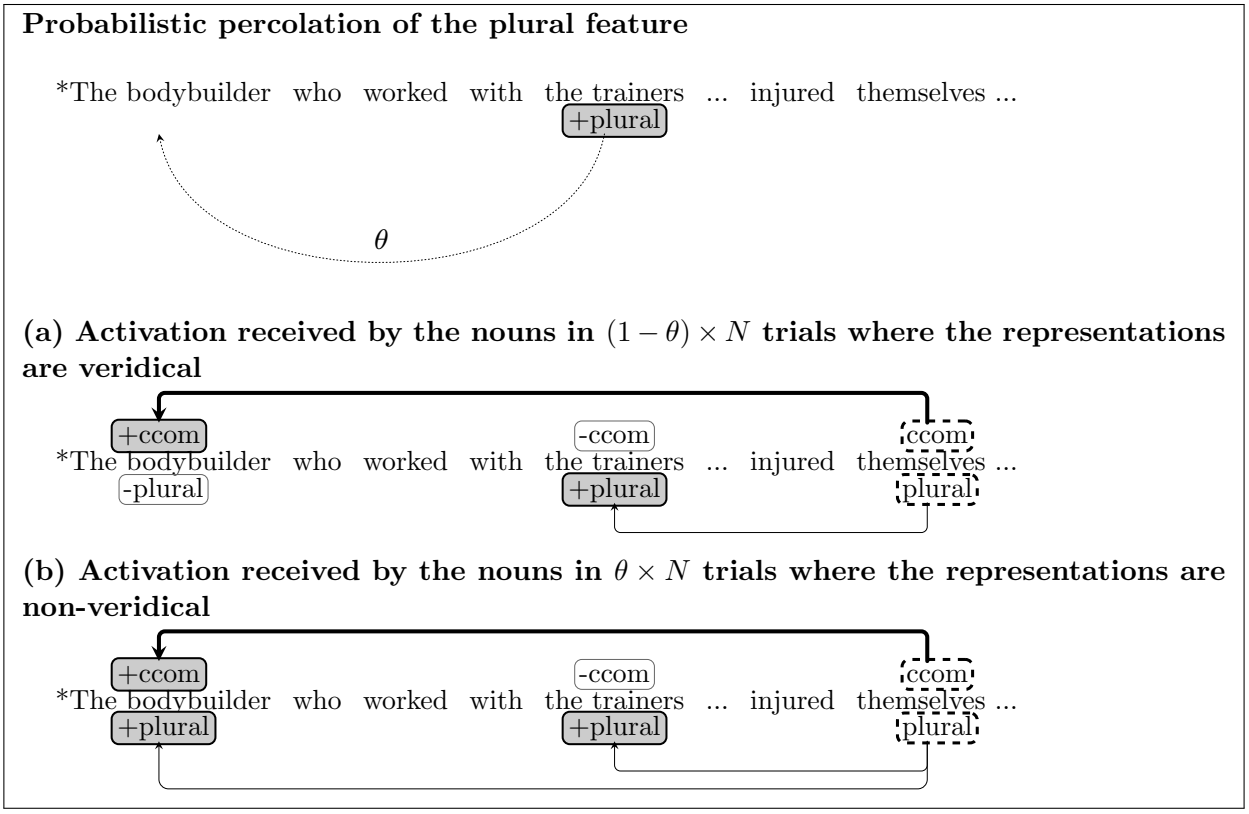


Figure 6.1: A densely dotted edge represents the probabilistic migration of a feature from one noun to another with probability θ. Activation received by the nouns based on cue-feature match is represented by solid arrows, with a thick arrow denoting more activation spreading compared to a thin arrow. A dashed box represents a retrieval cue, a shaded box represents a feature that matches a retrieval cue, and a thin box represents a feature that does not match a retrieval cue. The [ccom] and the [plural] cues together trigger a search in memory for a plural noun that c-commands the reflexive *themselves*.

to possess a plural feature in those trials. When the retrieval is triggered at *themselves*, multiple nouns (partially or fully) match the retrieval cues [c-command] and [plural]. All the matching nouns receive a certain amount of activation based on Equation 6.2. The noun with the highest activation gets retrieved and the retrieval time is determined by the activation level of the retrieved noun according to Equation 6.4.

6.1.3 The predictions of the distortion-retrieval model

A good model of sentence comprehension should be able to predict the processing difficulty exhibited by humans in resolving different kinds of dependencies. That is, the model-generated effects should be consistent with the effects observed in experimental data from humans on different dependency types. Some of the well-studied dependency types in this regard are subject-verb non-agreement dependencies, number agreement dependencies, antecedent-reflexive dependencies, and plausibility (mis)match configurations. I generate quantitative predictions from the distortion-retrieval model for each of these dependencies. Such quantitative predictions can be compared against future data to test if the distortion-retrieval theory can account for the effects observed for a particular dependency

type.

I use a Bayesian approach to generate model predictions: First, a prior distribution is specified for each free parameter of the model; this prior distribution could be based on our prior knowledge about a dependency or some reasonable assumptions about the plausible values of the parameter. Second, parameter values are repeatedly sampled from the priors and are used to simulate reading times, from which the effect of interest can be derived.

Subject-verb non-agreement dependencies

Consider the following pair of sentences:

(5) a. ... the resident who said that the neighbour was dangerous was complaining$_{subject}$ ...
b. ... the resident who was living near the dangerous neighbor was complaining$_{subject}$ ...

In both (5a) and (5b), when the verb phrase *was complaining* is encountered, the comprehender should figure out who was doing the act of *complaining*. Under the cue-based retrieval assumption, a search is triggered in memory for a subject noun phrase. In sentence (5a), both the nouns *the resident* and *the neighbour* are in subject position, i.e., they both match the retrieval cue [subject], compared to sentence (5b) where only one noun *the resident* matches the [subject] cue. The reading times at the verb *was complaining* are predicted to be slower in sentence (5a) compared to sentence (5b). This predicted effect is called *syntactic interference* (Van Dyke, 2007, Van Dyke and Lewis, 2003).

Similarly, when multiple nouns match the verb's semantic cues, such as [animate], they are assumed to cause *semantic interference* Van Dyke (2007). For example, in sentence (6c) where both *the resident* and *the neighbour* are animate, the retrieval times at the verb are predicted to be slower compared to sentence (6d), where only *the resident* is animate. The syntactic and semantic interference are the key predictions of the original cue-based retrieval model of Lewis and Vasishth (2005) and these predictions are supported by reading times data from English (Arnett and Wagers, 2017, Mertzen et al., 2022, Van Dyke, 2007, Van Dyke and Lewis, 2003, Van Dyke and McElree, 2011).

(6) a. ... the resident who said that the neighbour was dangerous was complaining$_{animate}$ ...
b. ... the resident who said that the warehouse was dangerous was complaining$_{animate}$ ...

What would the distortion-retrieval model predict for sentences in (5) and (6)? We first need to specify the prior distributions for the three parameters of the model, the latency factor, the feature distortion, and the cue-weighting. The latency factor F is the scaling parameter of the model and determines the range of reading times generated by the model. For example, a smaller value of F would produce faster reading times and a larger value of F would produce slower reading times. Thus, the parameter should be constrained in such a way that the model-generated reading times are neither too fast nor too slow compared to human reading times in sentence processing studies. As we discussed in Article I, the reading times are typically distributed such that their 2.5% quartile is greater than 150 milliseconds and their median is around 250–300 milliseconds. To match these properties, I choose the same prior on latency factor that we used in our articles.

$$F \sim Normal_{lb=0.05}(0.15, 0.05)$$

where $lb = 0.05$ represents the lower bound on values of the values of F. This lower bound is required to ensure that the model-generated reading times are not too fast compared to human reading times.

The second parameter, the distortion rate θ, determines the degree of feature distortion when the nouns are stored in memory, e.g., in what proportion of trials a feature migrates from one noun to the others. However, we need to specify what kind of features are allowed to be distorted when the nouns are stored in memory. For the current model, I assume that the features that are probably encoded in real-time during reading can undergo distortion, e.g., features like number, gender, case, etc. Other features like animacy, and lexical properties are unlikely to get distorted. Thus, for the non-agreement dependencies in 5 and 6, I assume that no feature distortion is possible; one can relax this assumption if the empirical data say otherwise.

$$\theta = 0$$

Finally, the cue-weighting encodes the relative weights of the cues that are used for retrieval. More specifically, it specifies the ratio of weights of a syntactic cue and a non-syntactic cue. For the subject-verb dependencies in (5) and (6), is the syntactic cue [subject] weighted higher than the semantic cue [animate]? This is possible! As we find in Article III, it is possible that the English speakers weigh the syntactic and semantic cues equally but the German speakers weigh the syntactic cue more strongly over the semantic cue in resolving subject-verb dependencies. There could be cross-linguistic differences in cue-weighting for these dependencies: In some languages, the syntactic and the semantic cues are weighted equally and in some languages, the syntactic cue is allowed to be weighted higher than the semantic cue. Thus, both assumptions should be considered for the cue weighting W; the data observed for a language can be generated by either of these two assumptions.

$$\begin{array}{rl} \textit{Equal cue weighting assumption:} & W = 1 \\ \textit{Stronger cue weighting assumption:} & W \sim Normal_{lb=1}(1, 1) \end{array}$$

I generate predictions from the distortion-retrieval model conditional on the above assumptions about cue weighting, distortion rate, and latency factor. Figure 6.2 shows the prior predictions of the model under the two sets of assumptions. Assumption set 1 specifies equal cue-weighting and assumption set 2 allows stronger weighting for the syntactic cue. Under the stronger cue-weighting assumption, the model predicts a smaller semantic interference in the range of 2 to 25 milliseconds. All other predictions are the same as the original cue-based retrieval model's predictions. The future data for sentences like (5) and (6) in a language can be compared against these predictions to infer whether the distortion-retrieval theory explains the subject-verb non-agreement processing.

Subject-verb agreement dependencies

Consider the following sentences containing a subject-verb dependency between *the key* and *was/were*. In sentences (7a) and (7b), the subject agrees in number feature with the verb, while in (7c) and (7d), the subject-verb number agreement is violated making these sentences ungrammatical. The sentences containing such subject-verb number agreement dependencies have been extensively studied in sentence processing.

(7) a. The key to the cabinet was rusty.

b. The key to the cabinets was rusty.

c. * The key to the cabinets were rusty.

d. * The key to the cabinet were rusty.

Assumption set 1
(Equal cue weighting)

Distortion rate: $\theta = 0$
Cue weighting: $W = 1$
Latency factor: $F \sim Normal_{lb=0.05}(0.15, 0.05)$

Assumption set 2
(Stronger cue weighting)

Distortion rate: $\theta = 0$
Cue weighting: $W \sim Normal_{lb=1}(1, 1)$
Latency factor: $F \sim Normal_{lb=0.05}(0.15, 0.05)$

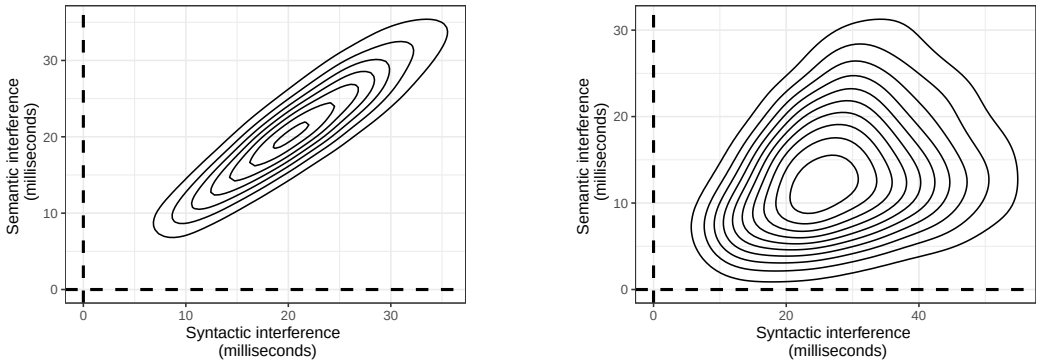


Figure 6.2: Syntactic and semantic interference effects (in milliseconds) predicted by the distortion-retrieval model for subject-verb non-agreement dependencies. The predicted effects are shown as a contour of the joint distribution of syntactic (on the x-axis) and semantic interference (on the y-axis). The left panel shows model predictions under the equal cue-weighting assumption and the right panel shows model predictions when the syntactic cue is allowed to be weighted higher than the semantic cue.

For the subject-verb number agreement dependencies, the predictions of the distortion-retrieval model have already been discussed in Article I.[3] The crucial assumption we made is that the plural feature of the non-subject noun *cabinets* in (7b) and (7c) can percolate to the subject noun and change its representation with a probability θ. This probability θ is the distortion rate. The following priors were set for the three parameters.

$$\begin{array}{rc} \text{Distortion rate:} & \theta \sim Normal_{lb=0.1}(0, 0.25) \\ \text{Cue weighting:} & W = 1 \\ \text{Latency factor:} & F \sim Normal_{lb=0.05}(0.15, 0.05) \end{array}$$

For the ungrammatical sentences, the model predicts *agreement attraction*: the reading times at the verb are predicted to be faster in (7c) compared to (7d). And, for the grammatical sentences, the model predicts that the effect of the number distractor can be positive, negative, or zero: the reading times in (7a) can be faster, slower, or comparable to (7b). I call this agreement distractor effect. Figure 6.3 shows the prediction space of the model for subject-verb number agreement effects. As we show in Article I, the data from 17 published studies on subject-verb number agreement are consistent with these predictions. Under the assumption that other agreement features such as gender can also get distorted, the model would predict the same attraction effects as shown in Figure 6.3 for any subject-verb agreement dependency.

[3]See the section "the feature-percolation-plus-retrieval model" in Article I (page 10).

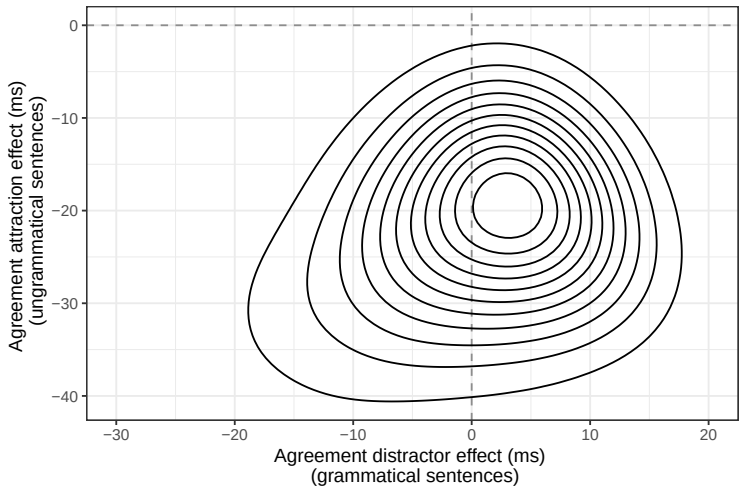


Figure 6.3: The agreement attraction and the agreement distractor effect (in milliseconds) predicted by the distortion-retrieval model for subject-verb agreement dependencies. The predicted effects are shown as a contour of the joint distribution of effects in the grammatical and ungrammatical sentences.

Dependencies involving semantic plausibility manipulation

Another kind of construction that has been used to investigate dependency completion process is the *plausibility (mis)match configuration* (Cunnings and Sturt, 2018). In these configurations, the semantic plausibility of the nouns is manipulated with respect to a verb; see sentences (8a-d). In sentences (8a) and (8b), the head noun phrase *the plate* is semantically compatible with its dependent relative clause verb *shattered*, making them semantic plausible sentences; but in (8c) and (8d), the noun phrase *the letter* is incompatible with the relative clause verb *shattered*, making them semantically implausible sentences. Under the cue-based retrieval account, the intervening noun *the cup* in (8a) and (8c) is predicted to cause an interference effect: In (8a), *the cup* would cause difficulty in retrieving the target head noun *the plate* as they both are semantic compatible with the verb *shattered*; in (8c), *the cup* would get occasionally retrieved at the verb *shattered* causing statistical facilitation compared to (8d), where neither of the nouns matches the verb semantically. Thus, the cue-based retrieval model of Lewis and Vasishth (2005) predicts — (i) *plausibility attraction effect*, a speedup in reading times at the verb in condition (8c) compared to (8d); and, (ii) *plausible distractor effect*: a slowdown at the verb in (8a) vs. (8b). The data from Cunnings and Sturt (2018) support these predictions.

(8) a. **Plausible Sentence, Plausible Distractor**
... Sue remembered the plate that the butler with the cup accidently shattered today ...

b. **Plausible Sentence, Implausible Distractor**
... Sue remembered the plate that the butler with the tie accidently shattered today ...

c. **Implausible Sentence, Plausible Attractor**
... Sue remembered the letter that the butler with the cup accidentaly shattered today ...

d. **Implausible Sentence, Implausible Attractor**
... Sue remembered the letter that the butler with the tie accidently shattered today ...

What are the predictions of the distortion-retrieval model for these configurations? Consistent with my assumption about other semantic features such as [+animate], I maintain that the semantic feature [+shatterable] is unlikely to get distorted when the nouns are stored in memory. Therefore, the distortion rate is assumed to be zero for the plausibility (mis)match configurations. However, regarding cue-weighting, the syntactic cue can be assumed have either equal or higher weight than the plausibility cue. Thus, two alternative assumptions can be made regarding cue weighting: (i) the syntactic and the semantic cues are weighted equal, or (ii) the syntactic cue can be weighted higher than the semantic cue. We get the same two sets of assumptions as we had for the subject-verb non-agreement dependencies.

	Distortion rate:	$\theta = 0$
Assumption set 1	Cue weighting:	$W = 1$
	Latency factor:	$F \sim Normal_{lb=0.05}(0.15, 0.05)$

	Distortion rate:	$\theta = 0$
Assumption set 2	Cue weighting:	$W \sim Normal_{lb=1}(1, 1)$
	Latency factor:	$F \sim Normal_{lb=0.05}(0.15, 0.05)$

Figure 6.4 shows the predictions of the distortion-retrieval model given the above prior assumptions. The only dataset on plausibility mismatch configurations (Cunnings and Sturt, 2018) supports the model predictions under the equal cue-weighting assumption. More experimental studies are needed to test these predictions of the distortion-retrieval theory.

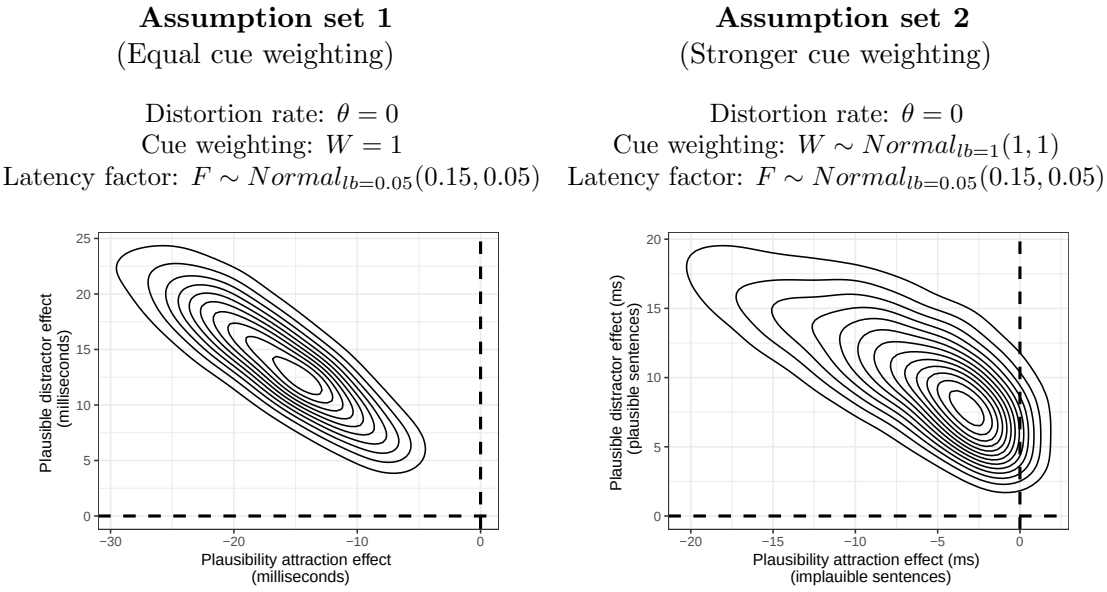


Figure 6.4: The plausibility attraction effect and the plausible distractor effect predicted by the distortion-retrieval model for plausibility (mis)match configurations. The predicted effects are shown as a contour of the joint distribution of effects in the plausible and the implausible sentences. The left panel shows model predictions under the equal cue-weighting assumption and the right panel shows model predictions when the syntactic cue is allowed to be weighted higher than the semantic cue.

Antecedent-reflexive dependencies

In sentences like (9a) and (9b), the reflexive pronoun *themselves* refers back to a noun phrase called its antecedent. The comprehender must identify the correct antecedent and link it with the reflexive. There is a hard syntactic constraint that plays a critical role in resolving the antecedent-reflexive dependency, *Principle A of the binding theory* (Chomsky, 1981). The principle states that an anaphor (e.g., a reflexive) must be bound within its governing category (e.g., its clause). Thus, in sentences (9a) and (9b), the antecedent would be a noun phrase that essentially c-commands the reflexive *themselves*. Under the cue-based retrieval assumption, a search is triggered in memory for a plural noun phrase that c-commands the reflexive *themselves*. Thus, the retrieval cues consist of a syntactic cue [c-command] and the number cue [plural]. In sentence (9a), *the bodybuilder* matches the c-command cue and *the trainers* matches the number cue. Consequently, a race for retrieval is initiated between the two nouns in (9a) compared to (9b) where only *the bodybuilder* matches a retrieval cue. This race process causes a statistical facilitation in retrieval times in sentence (9a) compared to (9b), i.e., the reading times at the reflexive are predicted to be faster in (9a) vs. (9b).

(9) a. The bodybuilder who worked with the trainers injured themselves ...

b. The bodybuilder who worked with the trainer injured themselves ...

However, a well-studied claim is that the c-command cue dominates the number cue in resolving antecedent-reflexive dependencies (Cunnings and Sturt, 2014, Dillon et al., 2013, Kush, 2013, Parker and Phillips, 2017). For example, Dillon et al. (2013) argued that the search for an antecedent of a reflexive is guided exclusively by Principle A of the binding theory implying that the number marking on the reflexive *themselves* is not used as a retrieval cue for these dependencies (also see Sturt, 2003). Several other researchers have hypothesized that the c-command cue may be weighted more strongly over number cues in processing reflexive dependencies (Cunnings and Sturt, 2014, Kush, 2013, Parker and Phillips, 2017). Our results in Article II support this cue weighting hypothesis. Specifically, we find that some individuals weigh the c-command cue higher than the number cue. Thus, the distortion-retrieval model should assume stronger weighting for the syntactic cue.

I generate predictions from the distortion-retrieval model under the following assumptions about its parameters.

$$\begin{array}{rc} \text{Distortion rate:} & \theta \sim Normal_{lb=0.1}(0, 0.25) \\ \text{Cue weighting:} & W \sim Normal_{lb=1}(1, 1) \\ \text{Latency factor:} & F \sim Normal_{lb=0.05}(0.15, 0.05) \end{array}$$

For the distortion rate, I have specified the same prior as in the case of subject-verb agreement dependencies. This is because the plural feature of *the trainers* can percolate to *the bodybuilder* in a proportion of trials; this number migration condition is the same as in subject-verb number agreement dependencies.

Figure 6.5 shows the prior predictions of the distortion-retrieval model given the above prior assumptions. Note that both representation distortion and differential cue-weighting components are contributing to these predictions. The percolation of plural feature from *the trainers* to *the bodybuilder* would enlarge the facilitatory effect in (9a) but the higher weights of the c-command cue attenuates the facilitatory effect to a larger extent causing an overall effect in the range of -29 to -3 milliseconds.

Table 6.1 summarizes the predictions of the distortion-retrieval model for different dependency types under different assumptions about cue weighting and distortion rate.

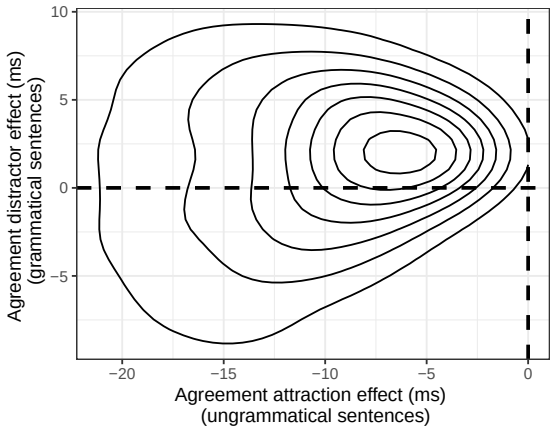


Figure 6.5: The agreement attraction and the number distractor effect (in milliseconds) predicted by the distortion-retrieval model for the antecedent-reflexive dependencies. The predicted effects are shown as a contour of the joint distribution of effects in the grammatical and ungrammatical sentences.

Table 6.1: **Predictions of the distortion-retrieval model for different dependency types:** the table shows the quantitative model predictions under each assumption for different dependency types. The square brackets represent 95% credible intervals of the predicted effect.

Dependency	**Assumptions**	**Predicted effects** (in milliseconds)	
		Syntactic interference	Semantic interference
Subject-verb non-agreement	Equal cue weighting	[9, 36]	[9, 36]
	Stronger cue weighting	[12, 51]	[4, 28]
		Plausibility attraction	Plausible distractor effect
Plausibility (mis)match		(implausible sentences)	(plausible sentences)
	Equal cue weighting	[-28, -7]	[6, 23]
	Stronger cue weighting	[-19, 0]	[4, 20]
		Agreement attraction	Agreement distractor effect
Subject-verb agreement	Probabilistic distortion,	(ungrammatical sentences)	(grammatical sentences)
	Equal cue weighting	[-40, -10]	[-20, 14]
		Agreement attraction	Agreement distractor effect
Antecedent-reflexive	Probabilistic distortion,	(ungrammatical sentences)	(grammatical sentences)
	Stronger cue weighting	[-29, -3]	[-13, 9]

6.2 Individual differences in sentence comprehension

Theories of sentence comprehension aim to describe the underlying cognitive process that uses linguistic knowledge to compute meaning out of a linear stream of words. To build these theories, one needs to infer the properties of the underlying process given a sample of individuals from the population. A conventional approach is to draw inferences from the average behavior of the sample (Kidd et al., 2018). The focus remains on the question of whether the processing behavior predicted by a theory is supported, or not, by the average behavior of the individuals in a sample. Even though the differences in processing behavior across individuals are commonly observed, this variability is either ignored or treated as a nuisance variable during modeling. For instance, in the case of mixed-effect regression models, the individual differences in model parameters are taken into account, but hypothesis testing still relies on the population-level estimates of the parameter(s) of interest.

The conventional approach of drawing inferences from the average behavior makes *the homogeneity assumption*: the cognitive process that underlies the observed behavior is invariant across individuals in a population. This assumption of homogeneity in the underlying cognitive process has been challenged by a number of methodological studies since the 1950s (see Estes, 1956, Hayes, 1953, Sidman, 1952, Underwood, 1975). For example, Estes (1956) demonstrated that the inferences drawn about the population from a curve fitted to grouped data may not generalize to the individuals in the population. While in many cases, the grouped behavior may reflect a great deal about the individuals' cognitive system, it would not hold in all cases. All these studies point out that individual-level behavior cannot be completely ignored in theory building.

The above argument finds considerable support in the data. Plenty of empirical studies in the last few decades have observed that individual differences are common and that the distribution of individual-level behavior may considerably deviate from the average behavior. The observations come from studies on perceptual decision making (Fific, 2014, Houpt et al., 2016), item recognition and free recall (Oberauer, 2005, Unsworth and Brewer, 2009, 2010), and psycholinguistics (Just and Carpenter, 1992, King and Just, 1991, MacDonald et al., 1992, McCauley and Christiansen, 2015, Pearlmutter and MacDonald, 1995), among others. Given these accumulating empirical data in favor of individual differences, it is hard to justify the conventional approach of focusing on the average behavior.

In sentence processing research, the individual differences gained traction in the 1990s when King and Just (1991) noted individual differences in processing object vs. subject relative clause in English and correlated these processing differences with the working memory capacity at the individual-level. This was followed by the Just and Carpenter (1992) study, where the authors theorized that working memory capacity differs across individuals, and the sentence comprehension behavior observed for an individual is determined by the individual's working memory score. Since then, many empirical studies have pointed out that individual differences do occur in sentence processing (see Farmer et al., 2012, for an overview) and the distribution of individual-level behavior may considerably diverge from the average behavior but the use of individual differences in theory building remains non-existent (see Kidd et al., 2018).

In the next two sections, I discuss the theoretical and empirical justifications for studying individual differences in sentence processing. After that, I discuss the challenges and future direction in drawing inferences from individual-level behavior.

6.2.1 Empirical reasons for studying individual differences

A growing number of studies have revealed individual differences in sentence processing (see Farmer et al., 2012, Kidd et al., 2018, for an overview). A commonly observed difference is that of quantitative differences in reading speed across individuals (Cheng et al., 2021, Cunnings and Fujita, 2021, Traxler et al., 2012). When the participants are asked to read sentences in an experiment, their average reading times per word differ from each other in a graded manner: some participants are faster than others. Figure 6.6 shows the distribution of individual-level reading speeds from a large-scale study on number agreement processing (Jäger et al., 2020). Here, reading speed is calculated as the reciprocal of the average reading time per word. The participants vary from a reading speed of 2 words per second to approximately 5 words per second (see Figure 6.6). Such differences in reading times, and by extension reading speed, have been shown to be highly *reliable* implying that a particular participant is likely to exhibit the same average reading speed across repeated experiments (Cunnings and Fujita, 2021, James et al., 2018).

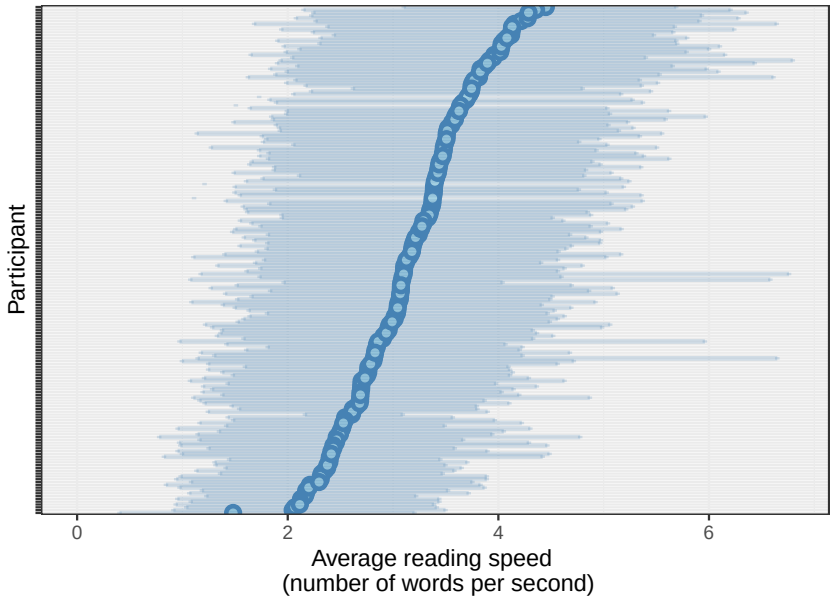


Figure 6.6: The distribution of individual-level reading speeds in the Jäger et al. study. The x-axis shows the average reading speed for a participant in the number of words per second; the average reading speed was calculated by taking the reciprocal of the average reading time (in seconds) per word. The y-axis represents each of the 181 participants in the study. The average reading speed across individuals varies between 2 words per second to approximately 5 words per second. The error bars show the uncertainty in the reading speed of an individual when it is measured across multiple sentences, possibly, driven by the processing difficulty level of the sentences.

It is not only the quantitative individual differences such as above, the qualitative differences in sentence processing have also been observed. In a recent study, Tanner (2019) observed that participants varied in their brain response on a continuum between N400 and P600 responses when they were asked to read subject-verb agreement violations like *The roses is ...* vs. *The roses are ....* Some participants showed N400 dominant response, while others showed P600 dominant response. However, the grand mean analysis of the brain response data suggested a typical left-negativity/P600 biphasic complexes. In such a situation, most of the individuals in the population

are not represented by the average behavior, thus, any inference based on average behavior may lead to inaccurate generalizations.

Many other studies in the last few decades have noted individual differences in sentence processing on a variety of different empirical phenomena including ambiguity resolution, pronoun resolution, and object vs. subject relative clause (e.g., Just and Carpenter, 1992, King and Just, 1991, MacDonald et al., 1992, Mätzig et al., 2018a, Novick et al., 2005, Pearlmutter and MacDonald, 1995, Swets et al., 2007). These processing differences have been attributed to varying working-memory capacity across individuals (Just and Carpenter, 1992, King and Just, 1991, Nieuwland and Van Berkum, 2006), or differences in language experience and learning (MacDonald and Christiansen, 2002, Misyak et al., 2010, Troyer and Kutas, 2020), or other sources such as cognitive control (January et al., 2009, Novick et al., 2005), perceptual processes (Dick et al., 2001, Leech et al., 2007), and chunking ability (McCauley and Christiansen, 2015).

Other comprehensive demonstrations of individual differences come from second language learners and from individuals with impaired sentence comprehension. In second language learning, it is observed that individuals differ in their language aptitude, motivation, learning strategy, and learning style (see Skehan, 1991, for an overview). This combination of learning factors impacts the development of second language in an individual (Dörnyei and Skehan, 2003), and consequently, produces individual differences in second language processing (Cheng et al., 2021, Hopp, 2015). In impaired comprehension studies, the individuals are shown to exhibit large variability in their performance on certain sentence processing tasks (Caplan et al., 2015, Mack et al., 2016, Mätzig et al., 2018b, Shammi et al., 1998). However, it is not entirely clear whether this variability is systematic or simply an artifact of noise, because, in a recent study, the observed rank of individuals with Aphasia in terms of their response time performance could not be replicated in a retest phase (Pregla et al., 2021).

In sum, a considerable number of studies have highlighted the occurrence of individual differences in sentence processing and have also indicated how these differences have implications for theory development.

6.2.2 Theoretical significance of drawing inferences from the individual-level behavior

When we focus on the average behavior of the participants to draw inferences, we make a critical assumption: The underlying cognitive process is homogeneous and the observed variability among individuals is due to the random noise in the homogeneous process (Levinson, 2012). This assumption implies that all the individuals in the population share the exactly same underlying system which is subject to measurement error in the experimental settings. This assumption is often either unstated or unjustified. While this assumption may be true in certain cases, there are other possibilities that remain under-explored. For example, it is possible that the observed individual-level differences are implicated by the systematic differences in the underlying cognitive processes across individuals. Several assumptions can be made about the linkage between the underlying cognitive process and the distribution of individual-level behavior in the population.

1. **Homogeneity assumption**: The individual-level behavior comes from a homogeneous process subject to noise, i.e., there are no differences in the underlying process across individuals. This assumption implies that a model with exactly the same set of mechanisms and the same set of parameter values should be able to correctly predict the behavior of all individuals on a processing task.

2. **Quantitative difference assumption**: The individual-level behavior comes from a heterogeneous process that has a continuous, graded distribution of parameter values across individuals in the population. According to this assumption, the model which allows graded variation in its parameter values can correctly predict individual-level behavior.

3. **Qualitative difference assumption**: The individual-level behavior comes from a heterogeneous process that either has a discontinuous distribution of the parameter values or a distinct set of underlying mechanisms across the individuals. This assumption implies that the underlying process can differ qualitatively across the individuals, either in terms of the range of parameter values or in the nature of underlying mechanisms. For example, one can assume that a certain parameter in the model takes a value 1 for most of the individuals and a value greater than 1 for some of the individuals, and these two parameter values have distinct theoretical implications. This approach allows us to constrain the source of individual-level variability in the model which aids in testing questions like whether are there any qualitative differences in the underlying cognitive processes across individuals.

Most of the studies in sentence processing rely on assumption (1), the *homogeneity assumption*. It could be the case that assumption (1) is correct in many cases, but this is an empirical question, which can be answered only by comparing the performance of the models under different assumptions. However, for that, we need to develop models under the assumption (2) and (3). The conventional approach of drawing inferences based on homogeneity assumption could posit several problems to theory development (Estes, 1956, Fific, 2014, Kidd et al., 2018, Maddox, 1999, Pachur et al., 2014).

First, the homogeneity assumption may lead to *inaccurate description* of the underlying cognitive process. This is because, in certain cases, the average behavior masks the theoretically important details about the underlying process (e.g., see Fific, 2014, Tanner, 2019). Consider the case of antecedent-reflexive dependencies, discussed in Article II. In sentences like *the bodybuilder who worked with the trainers injured themselves*, the cue-based retrieval theory predicts the facilitatory effect: the reading times at the reflexive *themselves* should be faster compared to a baseline sentence *the bodybuilder who worked with the trainer injured themselves*. However, Dillon et al. (2013) found no facilitation in these sentences which goes against the predictions of the cue-based retrieval theory. Their interpretations were based on the group-level reading times at the reflexive, not the individual-level data. In a large-scale replication of the same experiment, Jäger et al. (2020) did find a facilitatory effect based on the grouped data. The inferences based on average behavior in these two studies led to different, and possibly inaccurate, theoretical conclusions. In our Article II, we analyzed the individual-level data from these two studies. We found a strikingly similar pattern: in both studies, approximately three-fourth of the participants show a facilitatory effect and the remaining one-fourth do not show an effect. This result has an important theoretical implication: the reflexive dependencies are resolved via a cue-based retrieval process but individuals differ in how they weigh retrieval cues; only one-quarter of the population weighs the syntactic cue higher than the number cue in processing these dependencies. The models under the homogeneity assumption cannot produce such theoretical insight and they might produce an inaccurate description of the underlying process based on group-level data from a single study.

Second, the homogeneity assumption limits the prediction space of the model such that the model cannot capture the qualitatively distinct behavior across individuals on the same task. However, the researchers may not want too much degree of freedom in the model and they may intentionally limit the prediction space of the model; in certain cases, the qualitatively distinct behavior across individuals on the same task may contain theoretically important information.

Consider, the processing differences between healthy adults and Individuals with Aphasia (IWAs). Mätzig et al. (2018a) showed that a model of sentence processing implemented for healthy adults can also capture the qualitatively different comprehension behavior in IWAs by allowing individual-level variation in three parameters of the model. They showed that systematic variations in values of three parameters of the model are linked to individual-level processing behavior observed in healthy adults as well as IWA. Thus, the models of individual differences allow us to test theories that assume that the same set of underlying processes can generate qualitatively distinct behavior on an experimental task.

Third, the models under the homogeneity assumption may lead to *oversimplified description* of the underlying processes. Consider, the role of working memory constraints in sentence comprehension. The limited working memory resource has been implicated in determining processing difficulty in sentence comprehension (Gibson, 1998, Lewis and Vasishth, 2005). Just and Carpenter (1992) proposed that the differences in the working memory capacity lead to individual-level differences in performance on sentence processing tasks. This proposal opens up a window of possibility that working memory capacity could be a complex function in itself arising from different sub-processes which may vary across individuals. This exploration would not be possible if we make an oversimplified assumption that working memory capacity is homogeneous across the individuals in a population. Thus, the attempts to model individual differences based on assumptions (2) or (3) could benefit in developing a more specified description of underlying processes.

Finally, the models of individual differences based on assumption (2) could reveal theoretically important relationships among the components of the underlying system. Consider the hypothesis that during sentence comprehension, some individuals rely on predicting the upcoming linguistic material while others rely on minimizing working memory constraints. And, these two components of the comprehension system may interact to maximize the ease of the meaning computation process for an individual. Studying individual-level data is important for exploring such hypotheses. Thus, in situations where different components of the system may interact to optimize the functionality of the system, such a property of the system can be detected by modeling individual-level differences, which would otherwise be difficult to infer from the average behavior.

In summary, modeling individual differences is important for theory building because - (i) it allows a more accurate and more specified description of the underlying cognitive processes, (ii) it allows the model to capture qualitatively distinct behavior across populations using the same set of mechanisms, and (iii) it could reveal theoretically important relationships among different components of the underlying system.

6.2.3 Challenges and future direction

Using individual differences for theory development poses several challenges. A well-documented challenge is *the reliability paradox* (Hedge et al., 2018). When we interpret individual differences, we assume that the observed rank of participants in their scores on a task is not merely due to noise but reflects an underlying, true reality. The assumption implies that when we repeat the same task on the same set of participants, we should be able to rank the participants in the same order. The extent to which we can consistently rank the individuals across repeated tasks is called *reliability*. In a replication of seven classic cognitive tasks including Stroop task, Flanker task, etc., Hedge et al. (2018) found that most of these tasks produce surprisingly low reliability for individual differences. Similar observations were made by several other researchers: Individual differences exhibit low reliability especially when the between-subject variability is low or dominated by measurement errors such as trial-level noise (Cunnings and Fujita, 2021, Draheim et al., 2019, Hedge et al., 2018, Rouder and Haaf, 2019). These findings raised concerns because the same tasks

that produced robust effects at the level of average behavior were found unsuitable for measuring individual differences. Why do even the classical tasks produce such low reliability? Rouder et al. (2019) noted that the traditional experimental designs are incapable of separating the true between-subject variability from the trial-level noise. Due to a low number of trials per condition per subject in these tasks, the trial-level noise is large enough to mask the true individual differences. Given that the true individual-level variability could be quite low in certain scores, a reliable estimation of individual differences may require a very large number of trials, say 400 trials, per participant. Thus, the challenge of reliability emerges from a measurement challenge: we often do not have enough measurements from an individual to detect the true individual-level differences.

Another challenge comes from the modeling practices in sentence processing research. The computational models of sentence processing primarily focus on predicting the average behavior. When it comes to modeling the individual differences, a commonly taken approach is *the correlation-based approach*, where the correlation is computed between the hypothesized source of individual differences and the observed individual-level behavior (e.g., McCauley and Christiansen, 2015, Troyer and Kutas, 2020, Van Dyke et al., 2014). This approach provides a good starting point to identify the cognitive factors which may lead to individual-level variability in the processing behavior. But in order to use individual differences to draw direct inferences about the properties of the underlying cognitive process, one should attempt to model individual-level behavior in a computationally implemented model of sentence processing. A very few attempts have been made in this direction (e.g., Mätzig et al., 2018a).

The computational models need to be adapted to predict individual differences as a function of systematic variation in the parameter(s) or the structure of the model. An empirical bottleneck is in identifying which parameters of the model derive the observed differences in individual-level behavior. A two-step empirical test can be used to verify whether a particular parameter θ could be the source of individual differences: (i) Measure the parameter's value in each individual participant using a battery of tests and verify whether the parameter varies across individuals, (ii) Test whether the parameter value measured for an individual can predict the effect of interest for that individual on an independent task.

To summarize, both the above challenges arise because the experimental designs and the computational models have been developed for studying the average behavior. It would take a focused effort to develop the tools that are suited for drawing inferences from individual-level behavior. What approaches one can take to draw theoretical inferences from the individual-level behavior in a sample?

A sample of individuals from a population contains two types of information: (i) the distribution of individual-level behavior, which can answer how many participants show an effect consistent with model 1 and how many participants are consistent with model 2, and (ii) the behavior associated with each particular individual, which can answer whether a model predicts the effect observed for an individual X conditional on other measures taken from the same individual.

A method for testing theories using the first type of information, the distribution of individual-level behavior, has been demonstrated by Haaf and Rouder (2019). The authors propose that the models under different assumptions about the distribution of individual-level behavior can be used to draw inferences about the underlying process. Consider a model of the underlying process with a free parameter θ. We can make several assumptions about the distribution of the individuals in terms of parameter θ. Suppose, the first assumption is that all the individuals have positive θ but they differ quantitatively in its value; and suppose, the second assumption is that some individuals have positive θ and some have zero value of θ. Using participant-level data from an experiment, we can quantify evidence for each of these assumptions about the distribution of individuals in the population. We have used a similar approach in our Article III where a model assuming equal

weights for syntactic and semantic cues for all the individuals is compared against a model assuming varying cue weighting across individuals.

The second type of information, the behavior associated with each particular individual, is commonly used in psychometrics and, more generally, in the correlation-based approach to studying individual differences. In sentence processing research, the correlation-based approach has been used to link an individual's working memory capacity, reading experience, etc. with the individual's effect observed on an independent sentence processing task (McCauley and Christiansen, 2015, Troyer and Kutas, 2020, Van Dyke et al., 2014). However, the approach may not be useful in theory development because it oversimplifies the underlying relationship between individual-level processes and observed behavior. A more principled approach would be to directly model individual-level behavior as a variation in the model of the underlying processes. Under this approach, an individual will be measured on a battery of tasks in order to obtain independent measures of certain parameters such as working memory capacity and cue weighting; these individual-level parameter estimates will then be plugged into the model to generate predictions for that particular individual on a different experimental task. For example, one can first estimate the cue weighting and the reading speed for an individual and then use those estimates in a cue-based retrieval model to predict the magnitude of the agreement attraction effect for that individual; this predicted effect for the individual can be compared against observed attraction effect for the same individual on a number agreement processing task.

In sum, we can use a sample of individual-level behavior to infer: (i) the distribution of individuals in a population in terms of their underlying cognitive processes, and (ii) the properties of underlying processes associated with a particular individual.

The above approaches will help in developing more complete theories of sentence processing that can predict: (a) population-level behavior, (b) distribution of individual-level behavior, and (c) behavior of an individual on a task given other independent measures from the same individual. Such new approaches to theory building have implications for other areas of research in psychology and linguistics. The experimental and statistical methods in psychological sciences have primarily focused on whether inferences drawn from a sample are generalizable to the whole population or not. But, for the theoretical and empirical reasons I discussed here, one must ask whether the inferences drawn for a population are generalizable to an individual of the population or not. This question is important because we are ultimately interested in a theory of an individual's mind, not an abstract, ideal mind of the population. The study of individual differences will complement the average behavior approach in building a more accurate description of the underlying processes that govern processing behavior in humans.

Chapter 7

Conclusion

In this dissertation, I presented my theoretical and methodological contributions to the study of agreement attraction, a well-attested phenomenon in sentence comprehension. The main findings can be summarized as follows:

1. The feature distortion assumption — that the feature representation of nouns stored in memory can change with time — is necessary for explaining the number agreement effects in both grammatical and ungrammatical sentences.

2. A general account of dependency completion processes —the cue-based retrieval— alone is insufficient for explaining the agreement attraction data; the best fit is achieved by a hybrid model that combines cue-based retrieval and a feature distortion process.

3. Probabilistic feature distortion of nouns stored in memory modulates the content-addressable search for the target noun.

4. The absence of the number attraction effect in antecedent-reflexive dependencies is because of the stronger weighting of the syntactic cue over the number cue by some participants.

5. Individuals may differ in cue weighting: some, but not all, individuals weigh the syntactic cue higher than the number cue in resolving antecedent-reflexive dependencies.

6. Individual differences can reveal theoretically important details which may otherwise be masked by the average behavior. For instance, in two studies on number attraction in reflexive dependencies, the average effect produced different theoretical conclusions, but the distribution of individual-level effects is strikingly the same.

7. The absence of semantic interference effect in German is, possibly, because most German readers weigh the syntactic cue higher than the semantic cue.

8. Approximate Bayesian Computation — a likelihood-free Bayesian inference method — can be used to compare computational models of any underlying cognitive process without compromising the complexity of the models.

Overall, these findings advance our theoretical understanding of the dependency completion processes implicated in sentence comprehension. Three assumptions seem necessary for a complete theory of dependency completion: (i) the co-dependents stored in memory undergo probabilistic feature distortion, (ii) dependency completion is driven by a content-addressable search in memory, and (iii) the linguistic cues used for searching a co-dependent can be weighted differentially.

The work presented here is an important step in the direction of building a general architecture to explain all key phenomena observed in sentence comprehension. For example, a model incorporating the above three assumptions would be able to explain all number agreement effects, similarity-based interference, reflexive processing, and semantic attraction. Future work should evaluate these assumptions using benchmark data from other phenomena including garden path effects, local coherence, structural forgetting, etc.

This research direction is inspired by Allen Newell's vision of developing a "sufficient theory of a genuine slab of human behavior": a sufficiently specified and well-constrained computational model built for a complete analysis of a complex cognitive task (Newell, 1973). My work focuses on developing a sufficient theory of sentence comprehension that can generate constrained quantitative predictions for multiple phenomena observed in sentence processing tasks. And, for the reasons discussed in the previous chapter, I would also add the individual difference component to this research goal: A general theory of sentence comprehension should also be able to predict an individual's behavior on a task conditional on other independent measures, such as cue weighting, from the same individual. Such a theory should find independent support for its assumptions in the broader cognitive science literature, which would help in developing a single unified theory of cognition that can account for many phenomena across different tasks, as envisioned by Newell (1973). The modeling work presented in this dissertation is an important attempt in this direction.

Methodologically, this work lays a framework to build and evaluate complex computational models of sentence processing without compromising the complexity of the models. The algorithms I have developed here use a likelihood-free Bayesian approach for parameter estimation and model comparison. These methods are important because future research in sentence processing would inevitably go towards more complex process models, for example, the extended SWIFT model of eye-movement control and reading (Engbert et al., 2022, Rabe et al., 2021) and the self-organized parsing models (Smith et al., 2021, Smith and Vasishth, 2022), where the likelihood function is unknown or difficult to derive analytically. With the growing complexity of the models, it is useful to develop methods based on Bayesian inference as it allows us to estimate uncertainty in model parameters and model performance.

Bibliography

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4):1036–60.

Arnett, N. and Wagers, M. (2017). Subject encodings and retrieval interference. *Journal of Memory and Language*, 93:22–54.

Avetisyan, S., Lago, S., and Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, 112:104087.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., and Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. *Cognition*, 11(3):245–299.

Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. *Scientific reports*, 6(1):1–14.

Bays, P. M., Catalao, R. F., and Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of vision*, 9(10):7–7.

Bock, K. and Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, 8(1):57–99.

Bock, K. and Miller, C. A. (1991). Broken agreement. *Cognitive psychology*, 23(1):45–93.

Caplan, D., Michaud, J., and Hufford, R. (2015). Mechanisms underlying syntactic comprehension deficits in vascular aphasia: New evidence from self-paced listening. *Cognitive Neuropsychology*, 32(5):283–313.

Cheng, Y., Rothman, J., and Cunnings, I. (2021). Parsing preferences and individual differences in nonnative sentence processing: Evidence from eye movements. *Applied Psycholinguistics*, 42(1):129–151.

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

Cunnings, I. and Fujita, H. (2021). Quantifying individual differences in native and nonnative sentence processing. *Applied Psycholinguistics*, 42(3):579–599.

Cunnings, I. and Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, 75:117–139.

Cunnings, I. and Sturt, P. (2018). Retrieval interference and sentence interpretation. *Journal of Memory and Language*, 102:16–27.

Danker, J. F., Fincham, J. M., and Anderson, J. R. (2011). The neural correlates of competition during memory retrieval are modulated by attention to the cues. *Neuropsychologia*, 49(9):2427–2438.

Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., and Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological review*, 108(4):759.

Dillon, B. W., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69:85–103.

Dittmar, M., Abbot-Smith, K., Lieven, E., and Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child development*, 79(4):1152–1167.

Dörnyei, Z. and Skehan, P. (2003). 18 individual differences in second language learning. *The handbook of second language acquisition*, 589.

Draheim, C., Mashburn, C. A., Martin, J. D., and Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, 145(5):508.

Drenhaus, H., Saddy, D., and Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 145–165.

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language*, 36:147–164.

Eberhard, K. M., Cutting, J. C., and Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, 112(3):531–59.

Engbert, R., Rabe, M. M., Schwetlick, L., Seelig, S. A., Reich, S., and Vasishth, S. (2022). Data assimilation in dynamical cognitive science. *Trends in Cognitive Sciences*, 26:99–102.

Engelmann, F., Jäger, L. A., and Vasishth, S. (2019). The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science*, 43(12):e12800.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, 53(2):134.

Farmer, T. A., Misyak, J. B., and Christiansen, M. H. (2012). Individual differences in sentence processing. *Cambridge handbook of psycholinguistics*, pages 353–364.

Fific, M. (2014). Double jeopardy in inferring cognitive processes. *Frontiers in Psychology*, 5:1130.

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Gillund, G. and Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological review*, 91(1):1.

Haaf, J. M. and Rouder, J. N. (2019). Some do and some don't? accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, 26(3):772–789.

Hammerly, C., Staub, A., and Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive psychology*, 110:70–104.

Häussler, J. (2009). *The emergence of attraction errors during sentence comprehension*. PhD thesis, University of Konstanz, Germany.

Hayes, K. J. (1953). The backward curve: a method for the study of learning. *Psychological Review*, 60(4):269.

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, 50(3):1166–1186.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, 16(2):96–101.

Hopp, H. (2015). Individual differences in the second language processing of object–subject ambiguities. *Applied Psycholinguistics*, 36(2):129–173.

Houpt, J. W., Yang, C.-T., and Townsend, J. T. (2016). Modeling individual differences in perceptual decision making.

Jäger, L. A., Mertzen, D., Van Dyke, J. A., and Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111.

James, A. N., Fraundorf, S. H., Lee, E.-K., and Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of memory and language*, 102:155–181.

January, D., Trueswell, J. C., and Thompson-Schill, S. L. (2009). Co-localization of stroop and syntactic ambiguity resolution in broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, 21(12):2434–2444.

Just, M. A. and Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1):122–149.

Kidd, E., Donnelly, S., and Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, 22(2):154–169.

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5):580–602.

Kush, D. (2013). *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing*. Phd thesis, University of Maryland, College Park, MD.

Lago, S., Acuña Fariña, C., and Meseguer, E. (2021). The reading signatures of agreement attraction. *Open Mind*, pages 1–22.

Lago, S. and Felser, C. (2018). Agreement attraction in native and nonnative speakers of German. *Applied Psycholinguistics*, 39(3):619–647.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., and Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*, 82:133–149.

Laurinavichyute, A. and von der Malsburg, T. (2022). Semantic attraction in sentence comprehension. *Cognitive Science*, 46(2):e13086.

Leech, R., Aydelott, J., Symons, G., Carnevale, J., and Dick, F. (2007). The development of sentence interpretation: effects of perceptual, attentional and semantic interference. *Developmental science*, 10(6):794–813.

Levinson, S. C. (2012). The original sin of cognitive science. *Topics in cognitive science*, 4(3):396–403.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.

MacDonald, M. C. and Christiansen, M. H. (2002). Reassessing working memory: comment on just and carpenter (1992) and waters and caplan (1996).

MacDonald, M. C., Just, M. A., and Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive psychology*, 24(1):56–98.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676.

Mack, J. E., Wei, A. Z.-S., Gutierrez, S., and Thompson, C. K. (2016). Tracking sentence comprehension: Test-retest reliability in people with aphasia and unimpaired adults. *Journal of Neurolinguistics*, 40:98–111.

MacWhinney, B., Bates, E., and Kliegl, R. (1984). Cue validity and sentence interpretation in english, german, and italian. *Journal of verbal learning and verbal behavior*, 23(2):127–150.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & psychophysics*, 61(2):354–374.

Mann, J. (1982). Atmosphere or red herring? *Journal of General Psychology*, 106:159–163.

Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., and Burchert, F. (2018a). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, 10(1):161–174.

Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., and Burchert, F. (2018b). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, 10(1):161–174. Allen Newell Best Student-Led Paper Award at MathPsych/ICCM 2017.

McCauley, S. M. and Christiansen, M. H. (2015). Individual differences in chunking ability predict on-line sentence processing. In *CogSci*.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, 29(2):111–123.

McElree, B. (2003). Accessing recent events. *Psychology of Learning and Motivation*, 46:155–200.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48:67–91.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Mertzen, D., Paape, D., Dillon, B., Engbert, R., and Vasishth, S. (2022). Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data.

Misyak, J. B., Christiansen, M. H., and Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in psychology*, 1:31.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium.

Nieuwland, M. S. and Van Berkum, J. J. (2006). Individual differences and contextual bias in pronoun resolution: Evidence from erps. *Brain Research*, 1118(1):155–167.

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, 5(3):263–281.

Oberauer, K. (2005). Binding and inhibition in working memory: individual and age differences in short-term recognition. *Journal of experimental psychology. General*, 134(3):368—387.

Pachur, T., Hertwig, R., and Wolkewitz, R. (2014). The affect gap in risky choice: Affect-rich outcomes attenuate attention to probability information. *Decision*, 1(1):64.

Parker, D. and Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, 94:272–290.

Patson, N. D. and Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *Quarterly Journal of Experimental Psychology*, 69(5):950–971.

Pearlmutter, N. J. and MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of memory and language*, 34(4):521–542.

Pregla, D., Lissón, P., Vasishth, S., Burchert, F., and Stadie, N. (2021). Variability in sentence comprehension in aphasia in german. *Brain and Language*, 222:105008.

Raaijmakers, J. G. and Shiffrin, R. M. (1981). Search of associative memory. *Psychological review*, 88(2):93.

Rabe, M. M., Chandra, J., Krü"gel, A., Seelig, S. A., Vasishth, S., and Engbert, R. (2021). A Bayesian approach to dynamical modeling of eye-movement control in reading of normal, mirrored, and scrambled texts. *Psychological Review*, 28:803–823.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, 85(2):59.

Rouder, J., Kumar, A., and Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail.

Rouder, J. N. and Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic bulletin & review*, 26(2):452–467.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, pages 28–34. Institut für Deutsche Sprache.

Schäfer, R. and Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *LREC*, pages 486–493.

Schlueter, Z., Williams, A., and Lau, E. (2018). Exploring the abstractness of number retrieval cues in the computation of subject-verb agreement in comprehension. *Journal of Memory and Language*, 99:74–89.

Schneider, D. W. and Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology*, 64(3):127–160.

Scotti, P. S., Hong, Y., Golomb, J. D., and Leber, A. B. (2021). Statistical learning as a reference point for memory distortions: Swap and shift errors. *Attention, Perception, & Psychophysics*, 83(4):1652–1672.

Shammi, P., Bosman, E., and Stuss, D. T. (1998). Aging and variability in performance. *Aging, Neuropsychology, and Cognition*, 5(1):1–13.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological bulletin*, 49(3):263.

Skehan, P. (1991). Individual differences in second language learning. *Studies in second language acquisition*, 13(2):275–298.

Smith, G., Franck, J., and Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, 124:101356.

Smith, G. and Vasishth, S. (2022). A software toolkit for modeling human sentence parsing: An approach using continuous-time, discrete-state stochastic dynamical systems.

Sohn, M.-H., Anderson, J. R., Reder, L. M., and Goode, A. (2004). Differential fan effect and attentional focus. *Psychonomic Bulletin & Review*, 11(4):729–734.

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, 60(2):308–327.

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition*, 114(3):447–454.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48:542–562.

Swets, B., Desmet, T., Hambrick, D. Z., and Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General*, 136(1):64.

Tanner, D. (2019). Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. *Cortex*, 111:210–237.

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., and Jonathan, E. (2012). Individual differences in eye-movements during reading: Working memory and speed-of-processing effects. *Journal of eye movement research*, 5(1).

Troyer, M. and Kutas, M. (2020). Harry potter and the chamber of what?: the impact of what individuals know on word processing during reading. *Language, Cognition and Neuroscience*, 35(5):641–657.

Tucker, M. A., Idrissi, A., and Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology*, 6(347).

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, 30(2):128.

Unsworth, N. and Brewer, G. A. (2009). Examining the relationships among item recognition, source recognition, and recall from an individual differences perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6):1578.

Unsworth, N. and Brewer, G. A. (2010). Individual differences in false recall: A latent variable analysis. *Journal of Memory and Language*, 62(1):19–34.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):407–430.

Van Dyke, J. A., Johns, C. L., and Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, 131(3):373–403.

Van Dyke, J. A. and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49:285–316.

Van Dyke, J. A. and McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2):157–166.

Van Dyke, J. A. and McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, 65(3):247–263.

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4):685–712.

Wagers, M., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61:206–237.