Crosslinguistic evidence against interference from extra-sentential distractors

Daniela Mertzen
Department of Linguistics, University Library, University of Potsdam, Potsdam, Germany
Anna Laurinavichyute
Department of Linguistics, University of Potsdam, Potsdam, Germany
Brian W. Dillon
Department of Linguistics, University of Massachusetts Amherst, USA
Ralf Engbert
Department of Psychology, University of Potsdam, Potsdam, Germany
Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany

February 28, 2024

Data availability: The data and reproducible analysis code can be retrieved from https://doi.org/10.17605/OSF.IO/QRCMV.
Correspondence: Please send correspondence to mertzen@uni-potsdam.de.

Abstract

Cue-based retrieval theories of sentence processing posit that long-distance dependency formation is guided by a cue-based retrieval mechanism: dependents are retrieved via retrieval cues associated with a verb. When retrieval cues match multiple similar items in memory, this leads to cue-based retrieval interference. A landmark study by Van Dyke and McElree tested interference from sentence-external items: retrieval cues were manipulated to (mis-)match semantically similar items presented prior to a target dependency. The support for interference of this type is weak, and only comes from English object cleft constructions. Our study provides a cross-linguistic investigation of interference from sentence-external items: Three eyetracking studies in English, German and Russian tested interference in the online processing of filler-gap dependencies under varying task demands. A fourth study attempted to replicate the Van Dyke and McElree study using self-paced reading. Bayes factors analyses show cross-linguistic evidence against interference from sentence-external items. A broader implication from these data is that cue-based retrieval interference is driven by sentence-internal distracting items, suggesting that a cue-based search is restricted to the current linguistic context.

Keywords: sentence processing; similarity-based interference; cue-based retrieval; eye-tracking; Bayesian data analysis; task demands

Introduction

Sentence comprehension requires us to rapidly form dependencies between non-adjacent words. For example, in (1) a dependency needs to be established between the verb *complained* and its subject *the resident* (Van Dyke, 2007).

(1) *The resident* who was living near the dangerous neighbor *complained*.

To successfully integrate such structurally and temporally distant sentence elements, a working memory system is required that can store partially analyzed linguistic material. Specifying these memory mechanisms that subserve dependency formation is a key task in modeling the architecture of language processing (Gibson, 2000; Lewis, 2000).

One account of sentence processing, cue-based retrieval, specifies that non-adjacent dependencies are formed using a cue-based retrieval mechanism (e.g., Lewis and Vasishth, 2005; McElree, 2000; Van Dyke, 2007; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2011). Cue-based retrieval accounts broadly hold that retrieval processes are a key bottleneck in sentence comprehension: Sentence elements are encoded in memory, and these linguistic encodings are later reactivated, or retrieved from memory, when they are needed to support ongoing processing. In example (1), at the verb *complained* (the retrieval site), a memory query is launched to retrieve the sentence-initial noun phrase (NP) *the resident* from memory. This retrieval process relies on so-called retrieval cues to reactivate the target representation in memory. Cue-based theories generally assume that syntactic as well as semantic cues are used for retrieval. Simplifying somewhat, at the verb *complained*, the retrieval cues {*grammatical subject*} and {*animate*} might be used to retrieve the target NP *the resident* that carries the matching grammatical role and animacy features.

The cue-based retrieval process is assumed to lead to similarity-based interference: the retrieval of a target item can be impeded by other items in the sentence, called distractors, which have syntactic or semantic features that are similar to the features of the target item. In (1), *neighbor* is a distractor that shares the animacy feature with the target NP *the resident*. When a retrieval cue such as {*animate*} matches more than one item, this creates a cue overload. An overloaded cue impedes access to the target that has been encoded in working memory; this cue overload is assumed to cause similarity-based interference. These interference effects are reflected in increased processing times in reading studies.

Clear evidence for similarity-based interference from sentence-internal distractors

There is clear evidence that similarity-based interference affects sentence comprehension. Several reading studies have investigated interference from distractors that intervene between the retrieval target and the retrieval site (retroactive interference studies, e.g., Arnett & Wagers, 2017; Cunnings & Sturt, 2018; Dillon et al., 2013;

Jäger et al., 2020; Keshev & Meltzer-Asscher, 2019; Nicenboim et al., 2018; Schlueter et al., 2019; Thornton & MacDonald, 2003; Van Dyke, 2007; Van Dyke & Lewis, 2003) ; an example is shown in (1). Other studies report interference from within-sentence distractors that precede the critical dependency (proactive interference studies; e.g., Cunnings & Sturt, 2014; Koesterich et al., 2021; Parker & Phillips, 2017; Van Dyke & McElree, 2011; Wagers et al., 2009).

For both proactive and retroactive interference configurations, the same underlying retrieval mechanisms are assumed, i.e., shared features of target and distractor items lead to increased processing time (e.g., Lewis & Vasishth, 2005; McElree, 2000; Van Dyke & McElree, 2011). An important finding is that proactive interference may affect dependency formation to a lesser degree than retroactive interference (Jäger et al., 2017; Van Dyke & McElree, 2011). This implies that in language processing, a stronger interference effect occurs when the distractor intervenes between two co-dependents like subject and verb.

Here, we focus on one line of research that investigates a special case of proactive semantic interference: interference from sentence-*external* distractors that participants were asked to memorize prior to reading a target dependency (e.g., Fedorenko et al., 2006; Gordon et al., 2002; Van Dyke et al., 2014; Van Dyke & McElree, 2006). These studies suggest that sentence parsing can be disrupted even by linguistic items that occur outside the sentence.

Is there evidence for proactive interference from sentence-external distractors?

Although there exist published claims that interference caused by sentence-external distractors can increase linguistic dependency completion difficulty, surprisingly, all of these claims are based on statistically non-significant results. In at least one case the results have been found to be non-replicable.

For example, the proactive interference studies by Gordon et al. (2002) and Fedorenko et al. (2006) directly manipulated memory contents, using a dual-task paradigm that consisted of a word memorization task and a sentence reading task. In Gordon et al. (2002), participants were required to memorize either three descriptive nouns (*poet, cartoonist, voter*) or personal names (e.g., *Joel, Greg, Andy*) before reading a critical subject- or object cleft sentence (the factor Cleft type). These sentences had as their subject and direct object either descriptive NPs (e.g., *It was the dancer that liked the fireman/the fireman liked...*), or names (e.g., *It was Tony that liked Joey/Joey liked...*). When the sentence-external nouns matched sentence-internal NPs in semantic category (the factor Match), processing times were numerically larger in object vs. subject clefts. Importantly, however, at the critical region (*liked Joey/Joey liked* or *liked the fireman/the fireman liked*), the Match and Cleft type interaction did not reach significance: $F1(1, 55)=2.37$, $p = 0.13$, and $F2(1, 47)=1.74$, $p = 0.19$.

(Gordon et al., 2002, p. 429); $MinF'$(1,97=1.003, p-value=0.319).[1]

Similarly, Fedorenko et al. (2006) investigated the Gordon et al. (2002) design, using non-clefted subject- and object-extracted relative clause structures. Another change in the Fedorenko et al. design was that they used a memory load of either one or three nouns. Fedorenko and colleagues report that when three nouns had to be memorized, interference was greater for object than subject-extracted relative clauses. Even though Fedorenko et al. (2006) argued that high-load memory items increase processing difficulty in object vs. subject relatives, the critical interaction between relative clause type and memory load was not significant: $MinF'$(1,72)=3.34,p=0.07.[2] In addition, the experiment design of the Gordon et al. (2002) and the Fedorenko et al. (2006) study cannot test for the exact source of interference effects. Any potential interference effects could also be the result of encoding interference, that is, erroneous encoding of similar linguistic items in memory rather than cue overload at the retrieval site (Lewandowsky et al., 2008; Oberauer & Kliegl, 2006).

Another important proactive interference study–which we attempt to replicate in the present study–is by Van Dyke and McElree (2006). The authors adapted the Gordon et al. (2002) design to explicitly test for cue-based retrieval interference from semantically-similar, sentence-external distractors. To achieve this, they manipulated the retrieval cues at the verb, as shown in Table 1. Their memorization task had three animate nouns. This was followed by self-paced reading of object-cleft sentences. In the sentences, the critical verb (*sailed/fixed*) was manipulated such that the semantic retrieval cue ({*sailable*}, {*fixable*})[3] matched either only the target NP *the boat*, or matched the target NP as well as the memory nouns *table, sink, truck*. Cue-based retrieval accounts predict that, if the retrieval cues at the verb cannot uniquely seek out the target NP, this creates interference due to cue overload. Because {*fixable*} matches the target as well as the memory nouns, interference, reflected in a reading time slowdown, is expected at *fixed* compared to *sailed*. The design included a baseline comparison condition with no memory load (referred to as 'No load' hereafter); in this condition, the same sentences as the Memory load conditions are shown, but without the memory nouns. In the No load conditions, no significant reading time differences are expected. This predicted reading time pattern in Load and No load

[1] Significance of $MinF'$ is required to show that an effect can be generalized over participants as well as items (Clark, 1973). A widely held but mistaken belief is that $MinF'$ is unduly conservative (see the discussion in Raaijmakers et al., 1999); as Forster and Dickinson (1976) demonstrated, $MinF'$ is conservative in the highly unrealistic situation where between subject and between item variability is low. In psycholinguistics, between subject variability is generally quite high, and between item variability can also be quite high despite a counterbalanced Latin square design (e.g., see Vasishth et al., 2013).

[2] A complete analysis of the data from Fedorenko et al. (2006) appears in this blog post: https://vasishth-statistics.blogspot.com/2021/08/a-common-mistake-in-psychology-and_13.html

[3] We do not propose that the target NPs are encoded with the lexically specific features +*sailable* or +*fixable*. Rather, these can be viewed as placeholders for semantic cues. A principled approach to defining semantic cues is described in Smith and Vasishth (2020).

conditions would be reflected in an interaction between the factors *Memory load* and *Interference*.[4] Only cue-based retrieval accounts predict this interaction.

The presence of an interaction would be consistent with the prediction that interference has its source at retrieval, that is, semantic retrieval cues are used during online dependency formation, and these cues can become overloaded when there are several semantically similar items in memory.

Table 1
Example item (Van Dyke & McElree, 2006).

Memory load

table sink truck

No interference

It was *the boat* that the guy who lived by the sea *sailed* in two sunny days.

Interference

It was *the boat* that the guy who lived by the sea *fixed* in two sunny days.

Van Dyke and McElree (2006) reported the expected Load × Interference interaction, supporting the predictions of cue-based retrieval interference from sentence-external distractors. However, as in the other two studies mentioned above, the statistical evidence for this interaction effect was not convincing: there was a significant effect by participants and items: $F_1(1, 55) = 4.07$, $p < 0.04$, $F_2(1, 35) = 5.58$, $p < 0.02$, but crucially, the $MinF'$ statistic was non-significant: $MinF'(1, 90) = 2.35$, $p = 0.13$). A subsequent study with the same design and stimuli (Van Dyke et al., 2014) used linear mixed models instead of repeated measures ANOVA; at the critical region, the reported statistics even show an unexpected negative sign on the Load × Interference interaction (estimate: -10.2 ms, 95% CI [-40, 19.8], with a t-value of -0.66); the post-critical region showed the expected sign but was also not significant (estimate: 39.4, [-44.6, 123.4], t-value 0.929). The offline comprehension accuracy using logistic mixed effects regression also showed no evidence for an interaction (t=-1.452).

Revisiting proactive interference from sentence-external distractors: the problem with small-sample studies. The literature review shows that the evidence for cue-dependent retrieval interference from sentence-external materials in online sentence comprehension is at best suggestive, but statistically inconclusive: none of the above studies were able to find the crucial statistically significant interaction between load and interference (in order to argue for an interaction, one has to statistically demonstrate one, Nieuwenhuis et al., 2011).

[4]The factor labeled *Interference* is strictly speaking a misnomer here, at least for the No Load/Interference condition: No semantically similar nouns such as *table, sink, truck* are shown prior to the sentence with the target dependency boat-fixed that could potentially cause interference. However, in order to remain consistent with the original paper by Van Dyke and McElree (2006), we retain the term *Interference* for this factor.

One potential reason for the inconclusive interaction effects in Gordon et al. (2002), Fedorenko et al. (2006), Van Dyke and McElree (2006), and Van Dyke et al. (2014) is that similarity-based interference generally exhibits relatively subtle effects (e.g., Rabe et al., 2024). This is also evident from Jäger et al. (2017), which reports quantitative estimates based on a meta-analysis of published reading studies on similarity-based interference effects. For the type of subject–verb dependencies considered here, Jäger et al. (2017) estimated that the interference effect in reading studies ranges from 2 and 28 ms.[5] These effects are small compared to effects of word frequency and word length (e.g., Boston et al., 2008).

Small effects like those in similarity-based interference can either be highly inflated due to Type M/S error (Gelman & Carlin, 2014; Jäger et al., 2020; Vasishth et al., 2018), or remain undetected if tested with the participant sample sizes that are routinely used in experimental psycholinguistics (e.g., Jäger et al., 2017; Nicenboim et al., 2018). The standardly used participant sample sizes of 40-60 subjects (Vasishth, 2023; Vasishth & Gelman, 2021; Vasishth et al., 2022) lead to relatively low prospective statistical power (the probability of detecting an effect of a particular magnitude in a planned experiment) when testing for such small effects.

The interference effect reported in the Van Dyke and McElree (2006) study with sample size 56 is likely to be a Type M error-based inflated estimate: The critical interaction has a reported mean of 40 ms, with a 95% confidence interval spanning approximately 1 and 81 ms, which is quite a lot of uncertainty (for extensive discussion about uncertainty quantification as an important source of information, see Cumming, 2014; Jäger et al., 2020; Kruschke & Liddell, 2018; Vasishth, 2023; Vasishth & Gelman, 2021; Vasishth et al., 2018; Vasishth et al., 2022).

There exist several large-sample studies in the psycholinguistic literature that suggest that inflated estimates from underpowered studies may not be replicable. These studies investigated various psycholinguistic phenomena and were not able to replicate the original effects (e.g., Jäger et al., 2020; Nicenboim et al., 2020; Nieuwland et al., 2018; Vasishth et al., 2018). Given that low power can lead to misleading inferences, the current study aimed for a larger participant sample size in order to establish whether the the proactive interference effect can be detected.

Motivation for the present study

Our study investigated proactive interference cross-linguistically and under varying task demands. The primary goal of this study was to carry out a larger-sample test of proactive cue-based retrieval interference from sentence-external material on

[5]The meta-analysis included all published self-paced reading and eye-tracking (first-pass reading times) experiments from Van Dyke and Lewis (2003), Van Dyke and McElree (2006), Van Dyke (2007) and Van Dyke and McElree (2011). Of those studies, only the Van Dyke and McElree (2006) study used a dual-task paradigm with sentence-external distractors. The attentional demands of this task are quite different from a common reading task. However, all these studies did investigate subject-verb dependencies, and the effect sizes reported in these studies are similar.

sentence-internal dependency resolution (henceforth 'proactive interference'), using a dual-task paradigm (Van Dyke & McElree, 2006).

Studying proactive interference in such a paradigm is key in determining the role that cue-based retrieval interference plays in the processing of linguistic dependencies, when distractors are held in memory prior to encountering the linguistic dependency. Support for cue-overload due to such sentence-external distractors would point to a retrieval mechanism that is not constrained to the sentence context, but that erroneously considers extraneous linguistic items as retrieval targets. On the other hand, if our results show that sentence-external distractors do not cause retrieval interference, this would imply that interference in sentence processing is not simply driven by an unstructured bag of words in memory. Evidence against interference in such settings would show that the distractors must be contextually relevant: distractors must belong to the structured linguistic context to come into play as retrieval candidates.

The motivation for investigating proactive interference cross-linguistically. Our second goal was to determine whether the proactive interference effect can be detected cross-linguistically. Phenomena closely related to similarity-based interference in sentence comprehension have been studied in a variety of languages (e.g., Chinese: Jäger et al., 2015; Hebrew: Ness and Meltzer-Asscher, 2017; Hindi: Vasishth and Lewis, 2006; Spanish: Lago et al., 2015; German, Russian: Laurinavichyute et al., 2017). However, we are not aware of any cross-linguistic work on proactive interference from extra-sentential distractors. If cue-based parsing mechanisms are an integral part of human language processing, as posited by cue-based theories, then proactive interference from sentence-external memory items should be observable not only in English but cross-linguistically. An over-reliance on English is not limited to the phenomenon under investigation in this paper, it is a general problem in cognitive science. The focus on English has adversely affected the generalizability of theoretical claims relating to human cognitive processes (Blasi et al., 2022). Many studies on sentence processing phenomena from languages like Armenian, Czech, German, Hindi, Mandarin, Persian, and Spanish have shown different effects compared with English (Avetisyan et al., 2020; Bhatia & Dillon, 2022; Chromý et al., 2023; Dillon et al., 2016; Husain et al., 2021; Konieczny, 2000; Lacina & Chromý, 2022; Mitchell et al., 1990; Paape et al., 2021; Safavi et al., 2016).

In the context of proactive interference from sentence-external distractors, an important question that remains unanswered in the literature is whether languages with rich, unambiguous case marking also show proactive interference? One clear possibility is that rich case marking attenuates the interference effect; this could happen because case marking could help in identifying the right target for retrieval. An alternative possibility is of course that rich case marking does not attenuate interference, or at least there may be no evidence that it does (Avetisyan et al., 2020). If we find proactive interference in multiple languages, then one could be more confident about the cross-linguistic generalizability of cue-based retrieval interference

from sentence-external distractors.

The current study investigates proactive interference in English, as well as German and Russian. Both German and Russian have richer morphological marking on nouns, in the form of overt nominal case marking and gender, compared to English. The overt case marking may make items in memory more easily distinguishable. Specifically, case marking on the retrieval target could make the target more distinguishable from the distractors, and consequently reduce interference (Hartsuiker et al., 2003; Nicol and Antón-Méndez, 2009, but cf. Avetisyan et al., 2020). On the other hand, if case marking does not play a role, then proactive interference effects should be of the same magnitude across languages.

Motivation for investigating proactive interference under varying task demands. Although our main goal was to investigate proactive interference, our planned experiments also provided a unique opportunity to study another important issue regarding similarity-based interference effects in general which has not received much attention in the interference literature: Are interference effects observable if readers only superficially process the linguistic configurations under investigation? The similarity-based interference prediction is contingent on a simple but important assumption that is implicit in all cue-based retrieval theories. All models of retrieval assume that all syntactic dependencies are completely resolved during real-time processing. But this is not at all necessarily true. For example, the good-enough processing account assumes that syntactic dependencies are not always resolved, but can remain underspecified (Ferreira et al., 2002; Logačev & Vasishth, 2016a, 2016b; Sanford & Sturt, 2002; Swets et al., 2008; von der Malsburg & Vasishth, 2013). A question that has not been addressed in the interference literature is whether proactive interference effects are also observable if readers only superficially process the linguistic configurations under investigation. Our study explored this question.

There is some support for the hypothesis that a less demanding task (simple vs. more complex comprehension questions) will lead comprehenders to underspecify certain syntactic relations. So far, this has been shown only for relative clause attachment ambiguities (for English, see Swets et al., 2008; for German, see Logačev and Vasishth, 2016a).

Under complex task conditions where within-sentence dependencies have to be established, interference effects would be expected. If a simple task can induce superficial processing and readers do not establish all syntactic dependencies, that is, in a proportion of the trials the critical retrieval from memory does not occur, then it would be expected that proactive interference effects are smaller in magnitude, or the effects may disappear altogether.

The current study

To test cue-based retrieval theories' prediction of proactive interference, we implemented the dual-task paradigm (Van Dyke & McElree, 2006), using eye-tracking, in English, German and Russian. For each language, we tested two versions of the

experiment. In one version, items were followed by complex comprehension questions to induce deep processing. In another version, the same participants saw new items that were followed by simple comprehension questions, inducing superficial processing. Each participant saw the two experiment versions one to three weeks apart. A summary of the experiments is shown in Table 2.

Study language	**Experiment version**	**Tested subjects**	**Number of items**	**Factors 2×2 design**
English	Complex	74	40	Load, Interference
	Simple		40	Load, Interference
German	Complex	122	40	Load, Interference
	Simple		40	Load, Interference
Russian	Complex	120	40	Load, Interference
	Simple		40	Load, Interference

Table 2
Summary of the experiments testing proactive interference. For each language, depth of processing was manipulated through question complexity across two experiment versions. One version had complex comprehension questions (deep processing), and the other version had simple comprehension questions (superficial processing). The same participants saw both experiment versions (in randomized order) one to three weeks apart. Within each experiment version, we tested for the expected Load × Interference interaction.

Experimental design and materials

The Memory load × Interference manipulation. Our experiments used a 2 × 2 fully-crossed factorial design with the factors Memory load (*Load, No load*) and Interference (*No interference, Interference*), previously implemented by Van Dyke and McElree (2006). Both factors are within-subjects, within-items manipulations.

One difference between the Van Dyke and McElree (2006) study and ours is that Van Dyke and McElree (2006) tested object-cleft sentences whereas our study uses non-clefted structures with two embedded relative clauses.

In Van Dyke and McElree (2006), the clefted object retrieval targets are in linguistic focus. Here, focus describes the emphasis or prominence that is ascribed to certain sentence constituents by the syntactic structure (Chomsky, 1971). Psycholinguistic research has shown that items in focus have more distinctive memory representations (Birch & Rayner, 1997; Ward & Sturt, 2007). For our materials, we wanted to avoid increasing the prominence of the object noun through clefting because an increased prominence may reduce the magnitude of an interference effect (Engelmann et al., 2020). Thus, for our materials of the English, German and Russian eye-tracking experiments, we used sentence structures with target object NPs that are not in linguistic focus, expecting that non-clefted stimuli may increase the effect size, potentially leading to

higher statistical power. Complex non-clefted double-embedding structures allowed us to create items that are similar, and thus comparable, across the three languages.

However, in the Discussion section of this paper, we report an additional direct replication attempt (Experiment 4) of the original self-paced reading study by Van Dyke and McElree (2006); there, we use the same experimental object-cleft sentences as the original study, as shown in Table 1.

Table 3 shows English example items for the complex (3A) and the simple (3B) experiment version. All sentences had two embedded relative clauses, the outer relative clause being an object-relative clause, the most embedded relative clause being a subject-relative clause. Multiple center-embedded sentences can be read like a list of unrelated words and reduce their comprehensibility (e.g., see Miller, 1962; Miller & Chomsky, 1963). To avoid 'list-like' sentences (such as 'NP1 NP2 NP3 VP3 VP2 VP1'), and to facilitate comprehension by improving the phrasing of the sentences (Fodor et al., 2016), a prepositional or adverbial phrase was added before the matrix-clause verb.

In (3A), the critical dependency is between the relative clause verb *sailed/fixed* and its object NP *The boat*. The Load conditions present a list of three concrete, inanimate, singular nouns. In the non-interfering Load condition, the memory nouns *table, sink* and *truck* are not plausible objects of the critical verb (*sailed*). By contrast, in the interfering Load condition, they are plausible objects of the critical verb (*fixed*). The sentences in the No load conditions were identical to the sentences in the Load conditions. Here, no memory nouns were shown.

In our study, experimental items were followed by yes-or-no comprehension questions (with a 50:50 'yes-to-no' ratio). The complex experiment version used 40 experimental items with 90 filler items, and the simple version had 40 new experimental and 90 new filler items. German and Russian example sentences, and filler sentences are described in Appendix A.

Data Availability. All items, data and reproducible code can be retrieved from https://doi.org/10.17605/OSF.IO/QRCMV/.

Plausibility ratings of the stimuli. For all three languages, a plausibility rating task established that the target NP (e.g., *the boat*) was a plausible object of the RC verb (e.g., *sailed* and *fixed*) in the No interference and in the Interference conditions. The task also established that the distracting memory nouns (e.g., *table, sink, truck*) had a higher plausibility rating in the Interference conditions (with *fixed*) compared to the No interference conditions (with *sailed*).

In each language, plausibility was rated for all 80 items, 40 from the complex and 40 from the simple experiment version. For this rating task, eight conditions were created that combined the critical verb (e.g., *sailed* or *fixed*) with either the target (e.g., *the boat*) or one of the three distractor nouns (e.g., *table/sink/truck*). This resulted in the eight simple sentences in Example 2. Half of the items had a feminine personal subject pronoun, and half of the items a masculine one.

These experiments were run online, using Ibex Farm (https://spellout.net/

Table 3
English example items.

A) 'Complex' version:

Memory load: table sink truck

No interference

The boat that the guy who lived by the sea *sailed* in the morning was very old.

Interference

The boat that the guy who lived by the sea *fixed* in the morning was very old.

No memory load: — — —

No interference

The boat that the guy who lived by the sea *sailed* in the morning was very old.

Interference

The boat that the guy who lived by the sea *fixed* in the morning was very old.

'Complex' question: *'Did the guy live by the sea?'*

B) 'Simple' version:

Memory load: car scooter motorcycle

No interference

The plane that the pilot who returned from the Seychelles *landed* during the storm was pretty unreliable.

Interference

The plane that the pilot who returned from the Seychelles *crashed* during the storm was pretty unreliable.

No memory load: — — —

No interference

The plane that the pilot who returned from the Seychelles *landed* during the storm was pretty unreliable.

Interference

The plane that the pilot who returned from the Seychelles *crashed* during the storm was pretty unreliable.

'Simple' question: *'Was a pilot mentioned in this sentence?'*

ibexfarm/). Participants who did not take part in the main study were asked to rate the plausibility of the items on a scale from '1' (very implausible) to '7' (very plausible).

(2) a. He fixed the boat.
b. He fixed the table.
c. He fixed the sink.
d. He fixed the truck.
e. He sailed the boat.
f. He sailed the table.
g. He sailed the sink.
h. He sailed the truck.

It was expected that conditions a-e would receive higher plausibility ratings than f-h. Figure 1 shows the estimated probabilities for each of the seven rating choices in each condition. To estimate the probabilities, we fit Bayesian ordinal regression models in brms (Bürkner, 2017; Bürkner & Vuorre, 2019, see supplementary materials). Overall, the pattern is as expected: each language shows high plausibility ratings for a-e, i.e., the verb–target combinations (*fixed/sailed the boat*) and the verb–plausible distractor combinations (*fixed the table/sink/truck*). By contrast, lower plausibility ratings are observed in f-h, i.e., the verb–implausible distractor combination (*sailed the table/sink/truck*).

The depth of processing manipulation. Depth of processing was manipulated through comprehension question complexity in two experiment versions. In one version, the questions were relatively complex, while in the other version, the questions were simple. Depth of processing was implemented as a within-subjects, between-items manipulation.

Complex version. In this experiment version, complex comprehension questions induced deep processing. Here, *complex* refers to questions that required the reader to resolve the dependencies in the sentence. One half of the questions targeted the non-critical, most embedded relative clause (RC) (see Table 3A). The other half of the questions targeted the main clause subject-verb dependency, e.g., *'Was it the boat that was old?'*. While this avoids drawing particular attention to the critical dependency, a limitation of these question types is that accuracy results for the formation of the critical dependencies cannot be reported.

Simple version. In this version, simple questions induced superficial processing because they did not require participants to resolve within-sentence dependencies. These questions targeted non-critical NPs in the sentence (Table 3B).

English vs. German and Russian materials. An example of the English sentences was shown in Table 3. Appendix A shows the full examples of the German and Russian sentences.

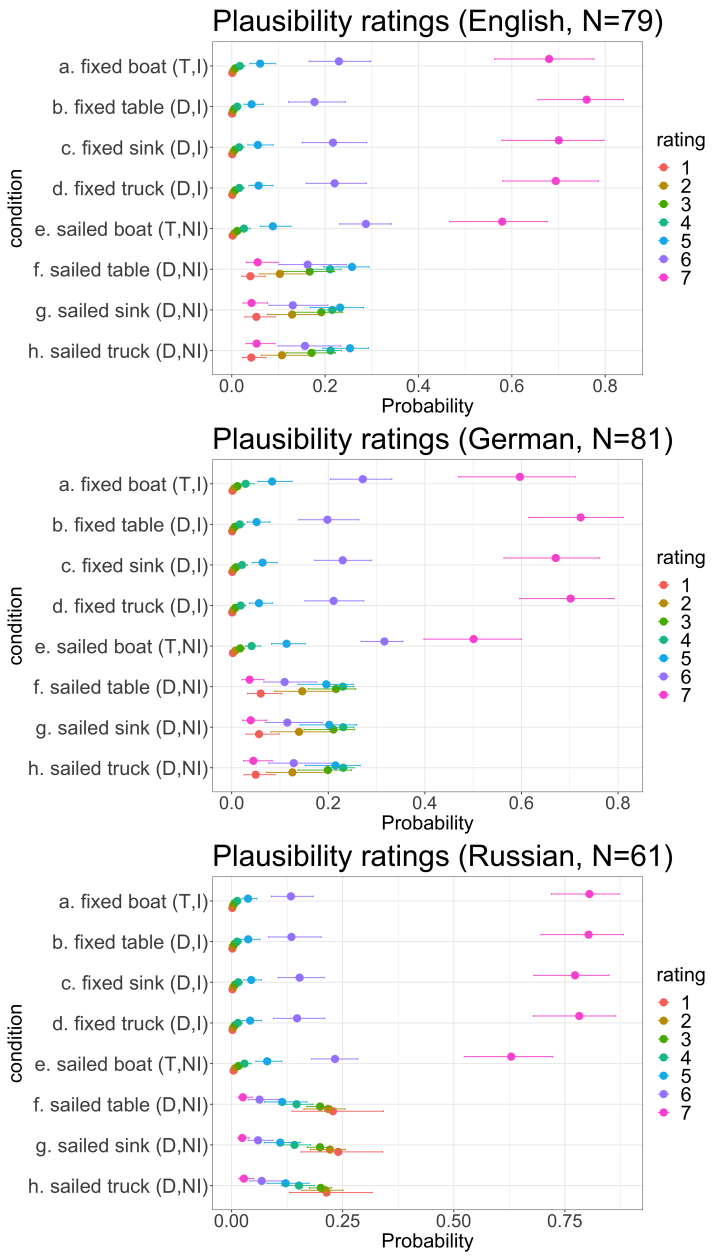

Figure 1. Plausibility ratings results for the items used in English, German and Russian. Shown are the estimates (posterior means with 95% credible intervals) of the probability of the plausibility rating in each condition. '1' = very implausible, '7' = very plausible. T = target, D = distractor, I = Interference, NI = No interference. For ease of interpretability, we use the English condition labels for all three languages.

Figure 2 shows example sentences (Load conditions) for all three languages schematically. For each language, in the non-interfering Load condition (a), only the sentence-initial target NP but not the memory nouns match the semantic cue of the verb (e.g., +*sailable*, +*drinkable*, or +*diagnosable*). In the interfering Load condition (b), the target NP as well as the memory nouns match the semantic cue at the verb (+*fixable*, +*smellable*, or +*discoverable*). In the interfering Load condition (b), this should result in an overloaded semantic cue.

In the English sentences, the target NP or the complementizer *that* are not overtly marked for case. By contrast, there is overt case marking on the determiner of the target NP in German and the demonstrative of the target NP in Russian as well as on the relative pronoun/complementizer that follows the target NP in both languages. The relative pronoun, which is the target of the object-retrieval, is marked for accusative case. If the relative pronoun is accessed for the object-retrieval, an additional assumption has to be made: that the semantic features are accessible at the relative pronoun. If case serves as an additional retrieval cue in German and Russian, this should lead to greater distinguishability of the retrieval target. More distinguishable items in memory may then lead to a smaller interference effect, compared with English.

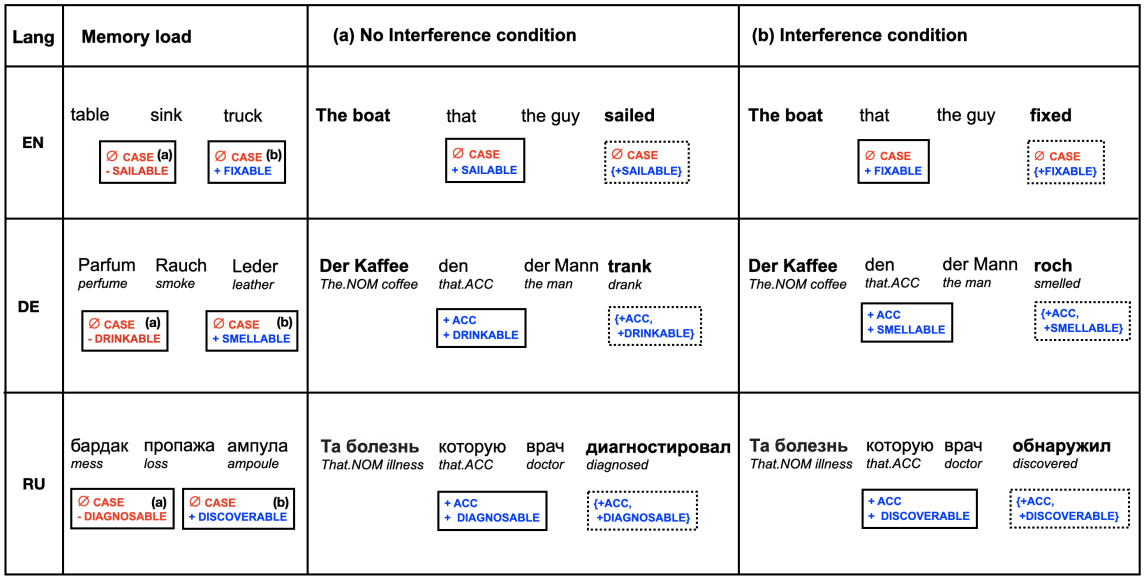

Figure 2. Example stimuli (schematic) for the English, German and Russian experiment.

Participants. All participants were native speakers of the tested language—English, German or Russian—with normal or corrected-to-normal vision and no history of language- or reading disorders.

An overview of the participant profiles is shown in Table 4. Some participants were not able to take part in the second part of the study. We excluded participants who did not have the tested language as their native language, who reported language

disorders, and participants with poor calibration due to pupil-detection loss. In addition, we excluded individual trials where the fixation sequence was disturbed by external influences.

The participant sample sizes resulted from collecting data until we reached a relatively precise uncertainty interval for the effect of interest (in total fixation times), i.e., an interval of width $\leq$ 40 ms (see the discussion on stopping rules in Kruschke, 2015; Vasishth et al., 2018). For English, due to pandemic-related logistical limitations, data collection stopped when we reached a 40 ms interval; if the pandemic had not intervened, we would have been able to reduce the width of the interval even further. For both German and Russian we obtained more precise estimates of the effect, because we were able to test up to 120 participants.

Study	**Analyzed subjects**	**Subject profile**	**Mean age (range)**	**Gender (%)**	**Recruitment; data collection location**	**Reimbursement**
English	66 (complex) 65 (simple)	undergrad	20 (18-27)	F: 66 M: 26 NB: 8	Linguistics & Psychological and Brain Sciences pool, posters; UMass Amherst, USA	30 USD or course credit
German	119 (complex) 122 (simple)	undergrad	25 (18-41)	F: 75 M: 25	Cognitive Sciences participant pool; Potsdam University, Germany	30 Euro or course credit
Russian	100 (complex) 109 (simple)	undergrad	22 (18-55)	F: 55 M: 45	Social media, word of mouth; Higher School of Economics, Moscow, Russia	course credit or interested volunteers

Table 4
Shown is a summary of the participant profiles for the English, German and Russian experiments. Undergrad means that participants were predominantly undergraduate students. F = female, M = male, NB = nonbinary.

Procedure. Each participants saw both experiment versions (complex or simple) that were shown in two sessions, seven to 21 days apart. The presentation order of the two versions was randomized.

In each session, after giving informed consent, participants read the study instructions which specified that both the reading and the recall task should be paid close attention to. The participants were seated in front of a presentation monitor, with their head in a chin- and head rest to minimize head movements. For monocular tracking of the right eye, we used a tower-mounted EyeLink 1000 (Plus) eye-tracker[6]

[6] https://www.sr-research.com/products/eyelink-1000-plus/

at a sampling rate of 1000 Hz. After an initial 9-point calibration- and validation procedure, each participant saw eight practice items. Experimental trials started after a further calibration.

The stimuli were presented according to a Latin Square design. Each of the four resulting lists contained one condition of each of the 40 experimental items interspersed with 90 filler items. The lists were randomized for each participant such that the items were not always shown in the same order. This was done to avoid that some items were always seen at the end of an experimental session when participants are fatigued. All sentences were displayed using a monospaced font in one line across the screen.

In each trial that showed a memory load, participants were presented with three memory nouns for a total of three seconds. Participants were asked to silently read and memorize the words.[7] Then, the memory nouns disappeared, and an experimental sentence was shown following a drift check. The 'drift check point' was located approximately at the same coordinates as the first letter of the first word in the sentence. Once participants finished reading a sentence, they fixated a small point in the lower right corner of the screen. This fixation triggered the presentation of the next screen. Both the location of the drift check point and the fixation trigger helped avoid random fixations on the sentence that are unrelated to reading. Finally, participants answered a question and were asked to recall the three memory nouns in the correct order, typing the answers.

The experiment had three hard-coded breaks to minimize fatigue, and to ensure relatively homogeneous study conditions for all participants. Re-calibrations were performed after each break, and whenever necessary. The differences in technical specifications for the English, German and Russian study can be found in the supplementary materials.

Predictions

For all three languages, an interference effect was predicted to occur at the outer RC verb (e.g., *fixed/sailed* in Table 3). This verb is the critical region in our experiment. Our primary analysis concerned the Load × Interference interaction, i.e., a reading time slowdown for interfering vs. non-interfering sentences within Load conditions, but not within No load conditions.

Predictions for complex vs. simple version. In the simple versions, it was expected that the Load × Interference interaction would be of a smaller magnitude than in the complex versions. An alternative hypothesis was that the effect may even disappear altogether when shallow processing is induced.

Predictions for the cross-linguistic comparison between German and Russian vs. English. In German and Russian, the magnitude of the Load ×

[7] In Van Dyke and McElree (2006), participants also saw the memory nouns at the beginning of a trial for three seconds but participants were requested to read them aloud. For our study, this change was necessary as the participants' head was placed in a headrest. Reading aloud would have required re-calibrations on most trials due to head movement.

Interference interaction could turn out to be smaller than in English. We reasoned that the richer morphological marking in German and Russian that may lead to better distinguishability of items in memory. If case can be used as an additional cue in German and Russian but not in English, this is expected to reduce interference from the sentence external distractor nouns in German and Russian.

Statistical analyses

We conducted our analyses within a Bayesian framework (Gelman et al., 2014). In the Bayesian setting, marginal posterior distributions can be computed which provide information about the plausible values of the parameters of interest. One assumption is that every parameter has a prior distribution of plausible values. The posterior can then be computed from the prior and a likelihood function, using Bayes' theorem (posterior $\propto$ prior $\times$ likelihood). In most cases—and this is true for the models we fit here—the posterior distribution cannot be derived analytically, but it can be approximated using Markov Chain Monte Carlo (MCMC) sampling (Gelman et al., 2014).

We fit Bayesian linear mixed-effects models, using the probabilistic programming language Stan (Carpenter et al., 2016). Memory load, Interference and their interaction were included in the models as fixed effects. The models had full variance-covariance matrices for by-subject and by-item random effects. In order to interpret the nested effects, we added centered word length as a predictor. This is because the manipulation of the critical verb region resulted in varying word lengths. The contrast coding for the comparisons of our statistical models is specified in Table 6. A log-normal likelihood was assumed for the reading times.

We used regularizing, weakly informative priors for the parameters in our models (Gelman et al., 2017). The prior distribution for the intercept was set at $\mathcal{N}(0, 10)$. All other parameters were defined as a $\mathcal{N}(0, 1)$ which, for the subjects and items random effects standard deviations, were truncated at 0. A so-called regularizing LKJ prior distribution was used for the correlation matrix associated with the variance-covariance matrix of the random effects (Lewandowski et al., 2009). Setting its shape parameter ν (nu) to 2.0 downweights extreme correlation values like ± 1.

For each of the statistical models, we ran four chains, each with 4000 iterations. The first half of these samples was discarded as warm-up, or burn-in, samples. The $\hat{R}$-diagnostic (Gelman et al., 2014) as well as visual inspection of trace plots were used to check for model convergence; all models reported here converged.

Bayes factor analyses for model comparison. The 95% credible intervals that we report for the parameters of interest in the linear mixed models are very informative because they give us estimates of the uncertainty of the effect given the model and the data (Cumming, 2014; Kruschke & Liddell, 2018). However, these intervals do not allow us to answer the question: is there evidence for the effect of interest? That question can only be answered by formal model comparison, which

essentially amounts to some version of a likelihood ratio test (Nicenboim et al., 2023; Royall, 1997; Vasishth et al., 2022).

In the Bayesian framework, the analog to the frequentist likelihood ratio test (aka ANOVA) is the Bayes factor (Gelman et al., 2014; Jeffreys, 1961; Kass & Raftery, 1995; Lee & Wagenmakers, 2014; Schad et al., 2022). The Bayes factor has several advantages over the frequentist analog, as discussed below. Accordingly, we conducted Bayes factor analyses to evaluate the evidence in favor of one model over another model. We explain the logic of the Bayes factor next.

In equation 1, Model 0 represents the null hypothesis which assumes the predicted interaction to be zero (the model does not include the interaction term). Model 1 represents the hypothesis that the interaction is not zero (the model includes the interaction term). Thus, the target parameter in Model 1 that represents the interaction is of primary interest. The Bayes factor (BF) is a ratio of the marginal likelihood of our data given one model over the marginal likelihood of the data given the other model:

$$BF_{01} = \frac{P(Data|Model_0)}{P(Data|Model_1)} \tag{1}$$

The word marginal in marginal likelihood is important: the Bayes factor is different from the frequentist ANOVA in that it computes the likelihood by summing together likelihoods, taking the uncertainty of the model parameter of interest into account (this is called integrating out the parameter). In the frequentist ANOVA, the likelihood is computed by plugging in the maximum likelihood estimate of the parameter of interest; as discussed above, this parameter estimate can be misleading if it is an overestimate (Gelman & Carlin, 2014).

In equation 1, the subscript 01 in BF_{01} shows which model is in the numerator and which one is in the denominator; here, the Bayes factor represents the evidence in favor of the null model M0 over the full model M1.

Provided that there is sufficient data (Vasishth et al., 2022), the Bayes factor can in principle tell us which of the two models is more likely to have generated the data. We thus have a way to quantify the support in favor of one model over another; even the null model can turn out to be supported (cf. the frequentist paradigm which is normally set up to only allow us to reject, or fail to reject, the null). Jeffreys (1961) as cited in Lee and Wagenmakers (2014) gives a guideline for the interpretation of the Bayes factor (Table 5, minimally adapted from Lee and Wagenmakers, 2014, p. 105).

An important feature of the Bayes factor is that it is sensitive to the prior distribution of the target parameter of interest (Gelman et al., 2017; Kass & Raftery, 1995; Schad et al., 2022; Sinharay & Stern, 2002). This is a great advantage over the frequentist likelihood ratio test because it allows us to evaluate the evidence for an effect of interest under varying prior assumptions/beliefs about the effect.

Mildly informative priors on the target parameter–these would be $\mathcal{N}(0, 1)$ on the log ms scale in our analyses–can strongly bias the Bayes factor in favor of the

Bayes factor (BF_{01})	**Interpretation**
> 100	Extreme evidence for M0
30 - 100	Very strong evidence for M0
10 - 30	Strong evidence for M0
3 - 10	Moderate evidence for M0
1 - 3	Anecdotal evidence for M0
1	No evidence
1/3 - 1	Anecdotal evidence for M1
1/10 - 1/3	Moderate evidence for M1
1/30 - 1/10	Strong evidence for M1
1/100 - 1/30	Very strong evidence for M1
$> 1/100$	Extreme evidence for M1

Table 5
Guidelines for the interpretation of the Bayes factor according to Jeffreys (1961) as cited in Lee & Wagenmakers (2014). The order of 0 and 1 in BF01 indicates that we look at support in favor of Model 0 over Model 1. BF10 indicates evidence for Model 1 over Model 0.

hypothesis that the target parameter is zero (e.g., Mulder & Wagenmakers, 2016; Rouder et al., 2018; Schad et al., 2022). We therefore computed the Bayes factor for more informative priors on the parameter representing the interaction term in Model 1: $\mathcal{N}(0, 0.1)$ and $\mathcal{N}(0, 0.05)$ (also see Nicenboim et al., 2020; Stone et al., 2023). These priors can be interpreted on the millisecond scale by back-transformation, but this requires knowledge of the mean reading time in the particular data being analyzed (Nicenboim et al., 2020). Given our mean reading times at the critical region (first-pass reading times or FPRT, and total fixation times or TFT) in the English, German, and Russian data, the relatively wider, less informative prior $\mathcal{N}(0, 0.1)$ implies that a priori, the interaction effect can lie (with 95% probability) between [-50, 50] ms in FPRT and [-80,80] ms in TFT. The more informative prior $\mathcal{N}(0, 0.05)$ implies that a priori, the effect can lie between [-25, 25] ms in FPRT and [-40, 40] ms in TFT. Given that the original estimate of the interaction reported in Van Dyke and McElree (2006) lies between [1, 81] ms, the wider prior $\mathcal{N}(0, 0.1)$ is the more appropriate one for the Bayes factor analysis. The more informative prior is included in the Bayes factor analysis because it is possible that the estimate from Van Dyke and McElree (2006) is an overestimate, due to Type M error.[8] Thus, the more informative prior would tell

[8]The transformation from log ms scale priors to milliseconds follows the approach in Nicenboim et al. (2023): given that the mean reading times at the critical region in the three studies range approximately from 5.5 log ms (FPRT) to 6 log ms (TFT), a $\mathcal{N}(0, 0.1)$ prior with ± 0.5 contrast-coded effects implies that the upper bound of the median first-pass reading times is $exp(5.5 + 0.2/2) - exp(5.5 - 0.2/2)$

us the evidence for or against the effect under an a priori belief that the effect size is relatively small; and the wider, less informative prior would yield evidence under the a priori belief that the effect is as large as in the original study by Van Dyke and McElree (2006). Both the priors allow the effect to be positive or negative in order to remain agnostic a priori about the direction of the effect—this ensures that the test does not optimistically assume that the interaction effect is positive in sign. We also consider a directional prior on the interaction that matches the estimated range of effects in Van Dyke and McElree (2006); this so-called enthusiastic prior (Spiegelhalter et al., 2004) would tell us whether we have evidence for the interaction, assuming a priori that the interaction is positive in sign.

The Bayes factors analyses were carried out by fitting models using the R package `brms`, an interface using Stan to fit Bayesian hierarchical models (Bürkner, 2017). For each model, we ran four chains with 80000 iterations (or 60000 iterations for the models using more informative priors) each. The first 5000 samples were discarded as warm-up samples. Marginal likelihoods and Bayes factors were computed using the bridge_sampler and bf functions from the `bridgesampling` R package (Gronau et al., 2017; Gronau et al., 2020).

Condition	Load	Interference	Interaction
a. Load, No interference	+0.5	−0.5	−0.25
b. Load, Interference	+0.5	+0.5	+0.25
c. No load, No interference	−0.5	−0.5	+0.25
d. No load, Interference	−0.5	+0.5	−0.25

Table 6
Contrast coding for effects of Load, Interference and their interaction.

Results

Comprehension question accuracy. The by-condition question response accuracies for all six experiment versions are shown in Figure 3.[9] In Van Dyke and McElree (2006), the by-condition accuracies are not reported, but the Load conditions overall had a lower accuracy than the No load conditions (95% CIs [77, 89]% vs. [81,

(FPRT), and $exp(5.5 + 0.2/2) - exp(5.5 - 0.2/2)$ (TFT); the lower bounds are simply these values with a negative sign. If the mean reading time is assumed to be 6 log ms (e.g., for total fixation time), the 5.5 is simply replaced with 6 in the above calculations.

[9]Statistical analyses of the question response accuracies are not shown here because the comprehension questions did not target the critical dependency. It would therefore be surprising to see an effect in the response accuracies. Indeed, the 95% CrIs of the interaction terms are centered on zero across all languages and versions. The analyses can be inspected in the supplementary materials.

93]%) and the Interference conditions had a lower accuracy than the No Interference conditions (95% CIs [77, 89]% vs. [81, 93]%).

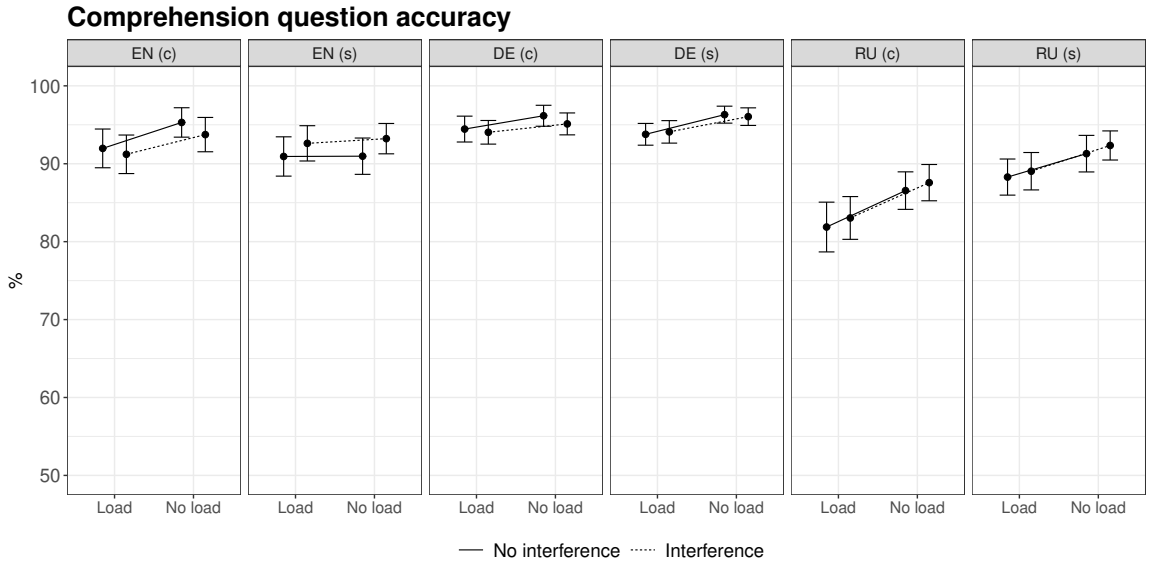

Figure 3. By-condition mean comprehension question accuracies (in percent) with 95% confidence intervals. EN = English, DE = German, RU = Russian; (c) = complex version, (s) = simple version.

Memory recall accuracy. Figure 4 shows the recall accuracy of the memory items for non-interfering vs. interfering Load conditions for each experiment version next to the Van Dyke and McElree (2006) recall accuracies. The accuracies in Figure 4A are based on a strict criterion where recall was judged as 'correct' only when all three memory nouns were recalled in the correct order.[10] Compared to Van Dyke and McElree (2006), our recall accuracies are low, particularly in English when complex comprehension questions were asked. To check that participants did not largely disregard this task, we inspected a more lenient criterion: recall accuracy was judged as 'correct' when either two or three memory nouns were recalled in any order. These results are presented in Figure 4B.

The lower recall accuracy in our study may be the result of the participants silently reading the nouns in the memorization task. In Van Dyke and McElree (2006), participants read the memory nouns aloud, possibly facilitating recall (MacLeod et al., 2010; Quinlan & Taylor, 2013). Another hypothesis is that our participants paid less attention to the recall task than the reading task (see Discussion).

Reading times. We report the results at the critical verb for the reading measures first-pass reading times (FPRT) and total fixation time (TFT). FPRT, also

[10] A more lenient criterion in Van Dyke and McElree (2006), i.e., recall of three words in any order, had a highly similar accuracy to the strict criterion (non-interfering 80%, interfering 78% (SE 2). Removing the strict order criterion also did not change the results for our data.

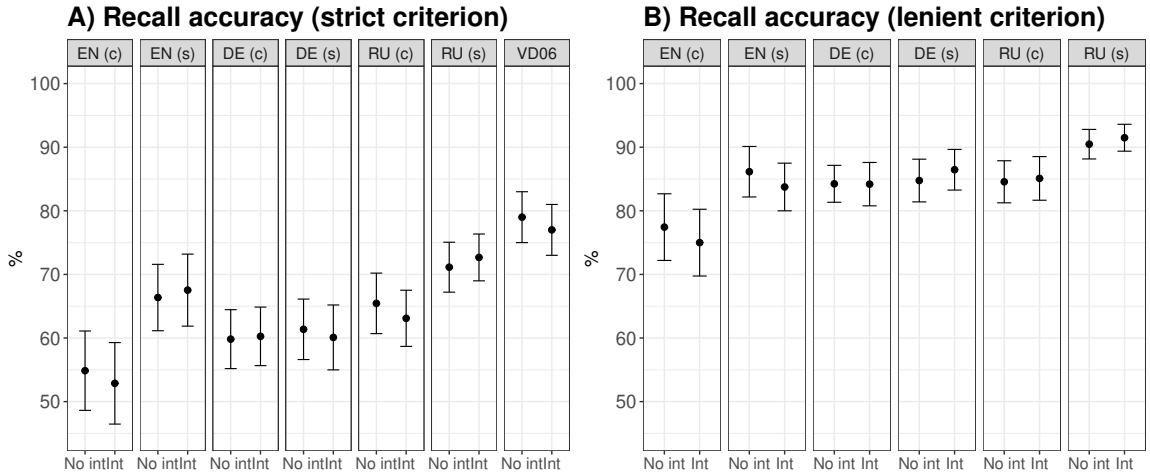

Figure 4. Mean recall accuracies (in percent) with 95% confidence intervals, A) for a strict criterion (all three words recalled in the correct order), and B) for a lenient criterion (two or three words recalled in any order). EN = English, DE = German, RU = Russian; (c) = complex version, (s) = simple version; VD06 = Van Dyke & McElree (2006); No int = No interference condition, Int = Interference condition.

referred to as gaze duration, is the sum of all fixations on a word n before any other word is fixated. TFT includes all fixations on a word n (Logačev & Vasishth, 2013; Rayner, 1998). The raw by-condition means with 95% confidence intervals for both measures can be inspected in Figure 5.

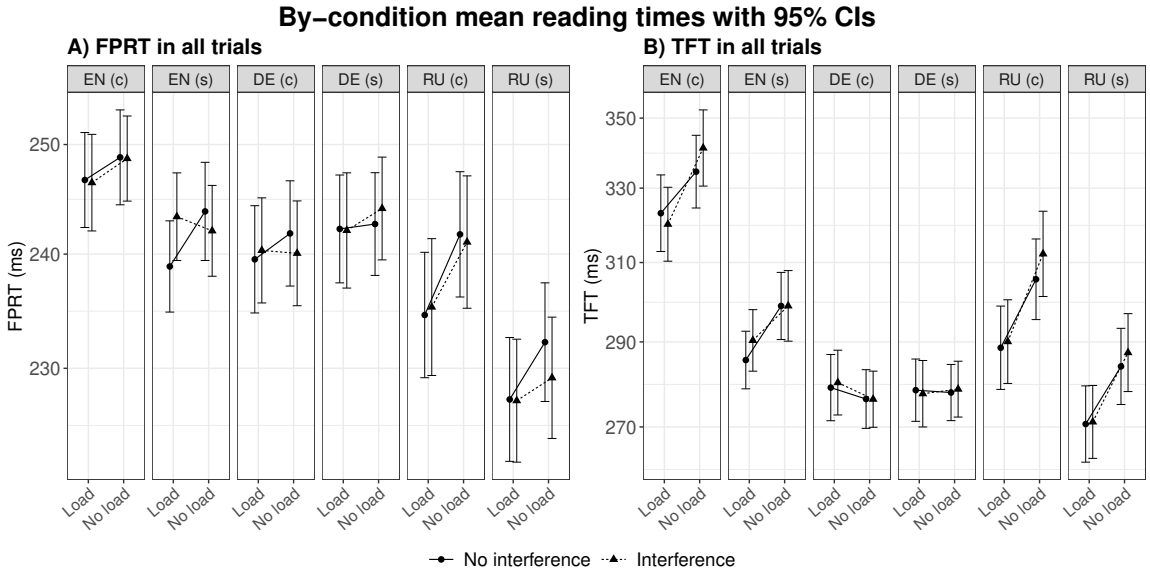

Figure 5. By-condition mean reading times with 95% confidence intervals. (A) shows first pass reading times (FPRT); (B) shows total reading times (TFT). EN = English, DE = German, RU = Russian; (c) = complex version, (s) = simple version.

Figure 6 shows the results of the Bayesian analysis for English, German and Russian, each in the complex version (left panel) and the simple version (right panel). For the effects of Load, Interference and their interaction, we show the FPRT and TFT means of the posterior distributions with their 95% credible intervals.

The analysis indicated that reading is overall faster when items have to be memorized and recalled, i.e., Load conditions were read faster than No load conditions. Such reading time speedups are regularly observed under a high processing load (e.g., Laurinavichyute et al., 2017; Nicenboim et al., 2018; Van Dyke & McElree, 2006). Meanwhile, there is no indication of an overall effect of Interference. This was expected because in this experiment design, once memory load is removed, any potential interference from the memory items is also removed. Therefore, the effect of Interference only tests whether there is a difference between the verbs in conditions a, c vs. b, d.

Effect estimates of the Load × Interference interaction. The effect of interest is the Load × Interference interaction. This interaction was expected to have a positive sign. Across all experiments, only the simple versions of English and Russian, in first-pass reading times, show some indication of an interaction in the expected direction. In English (simple, FPRT), the 95% CrI is [0, 37] ms. Nested comparisons show that within Load conditions, there is a reading time slowdown for interfering compared to non-interfering sentences (95% CrI [-1, 27] ms). In No load conditions, the 95% CrI is centered around zero ([-19, 9] ms). In Russian (simple, FPRT), the interaction points in the expected direction (95% CrI [-7, 26] ms). In Figure 7, our FPRT interaction estimates from each experiment are presented side by side with the Van Dyke and McElree (2006) interaction. Both English and Russian show effect estimates that overlap with the interaction in Van Dyke and McElree (2006).

In contrast to English and Russian, the simple version of German shows a pattern that is inconsistent with the expected Load × Interference interaction (FPRT: 95% CrI [-23, 2] ms, nested Load: [-15, 12] ms, nested No load: [-4, 21] ms). For TFT, the 95% CrI is [-32, 0] ms (nested Load: [-25, 9] ms, nested No load: [-11, 27] ms). Similarly, Russian (complex) also shows a negative sign for the interaction (FPRT: 95% CrI [-34, -2] ms; nested Load: [-30, 0] ms, nested No load: [-9, 18] ms; TFT: 95% CrI [-53, 3] ms; nested Load: [-21, 29] ms, nested No Load: [1 to 63] ms). For all other experiment versions, the Load × Interference interaction is centered around zero.

Bayes factor results for the Load × Interference interaction. The Bayes factor results can tell us whether there is evidence in favor of or against the Load × Interference interaction. Figure 8 shows the Bayes factor results (BF_{01}) for Model 0 (not including the interaction term) over Model 1 (including the interaction term), separately for the complex and simple versions in each language. Panel (a) visualizes results for first-pass reading times, and (b) for total reading times. The Bayes factor values (y-axis) were computed using prior distributions centered on 0 and with increasingly uninformative standard deviations (SD) on the interaction (x-axis).

Referring to the guidelines in Table 5 for interpreting the Bayes factors, we can

see in Figure 8 that for almost all experiment versions in FPRT and TFT, the Bayes factors show either no evidence or suggest that the data are more likely to have been generated by Model 0, that is, showing anecdotal to moderate evidence against a Load × Interference interaction. Only for English (simple, FPRT), the Bayes factor very weakly supports the expected interaction but only under a more informative prior. For Russian (complex, FPRT), the Bayes factor very weakly supports an interaction if we use a more informative prior; however, crucially, this interaction is in the *unexpected* direction. Thus, with the exception of English (simple, FPRT), there is either no evidence at all from all three languages for the expected sign of the interaction, or there is evidence against the expected sign of the interaction.

The most optimistic Bayes factor analysis we could do is to use the estimates of the interaction from the original Van Dyke and McElree study as priors–this is the enthusiastic prior analysis. Because the prior for the interaction is based on estimates from self-paced reading data, we only analyzed total reading times as they more closely approximate self-paced reading data (first-pass reading times would exclude any later-stage processing, and tend to be much smaller in magnitude than self-paced reading estimates). Even using enthusiastic priors on total reading times show (Figure 9) that the Bayes factors either furnish no evidence for the interaction, or show anecdotal to moderate evidence against the expected interaction effect.

Effect estimates of the Language × Load × Interference interaction. Although we did not observe the predicted Load × Interference interaction, for completeness we further tested whether the Load × Interference interaction is modulated by language. The right panel of Figure 10 shows that the interaction has a positive sign for FPRT as well as TFT. The three-way Language × Load × Interference interaction indicates that processing in English may differ from German and Russian.

In addition to the Language manipulation, we manipulated depth of processing. This manipulation did not show the expected pattern, namely, a smaller, or no interference effect in the simple compared with the complex experiment versions. In English, the reading time patterns are compatible with the idea that processing differs across the two versions. However, unexpectedly, interference is observed only in the simple version (for Processing depth × Load × Interference interactions).

Discussion

This study investigated proactive cue-based retrieval interference in sentence comprehension. Overall, our large-sample cross-linguistic study either does not show any evidence, or shows evidence against the predicted interaction. The only exception to this broad generalization was seen in the English group's first-pass reading times in the shallow testing condition; here, a Bayes factor analysis very weakly supported the expected interaction under a more informative prior.

If this is an accurate generalization about the contexts where proactive interference from extra-sentential memory lists occurs, then this finding limits the scope of cue-based retrieval interference in sentence comprehension. But before drawing any such

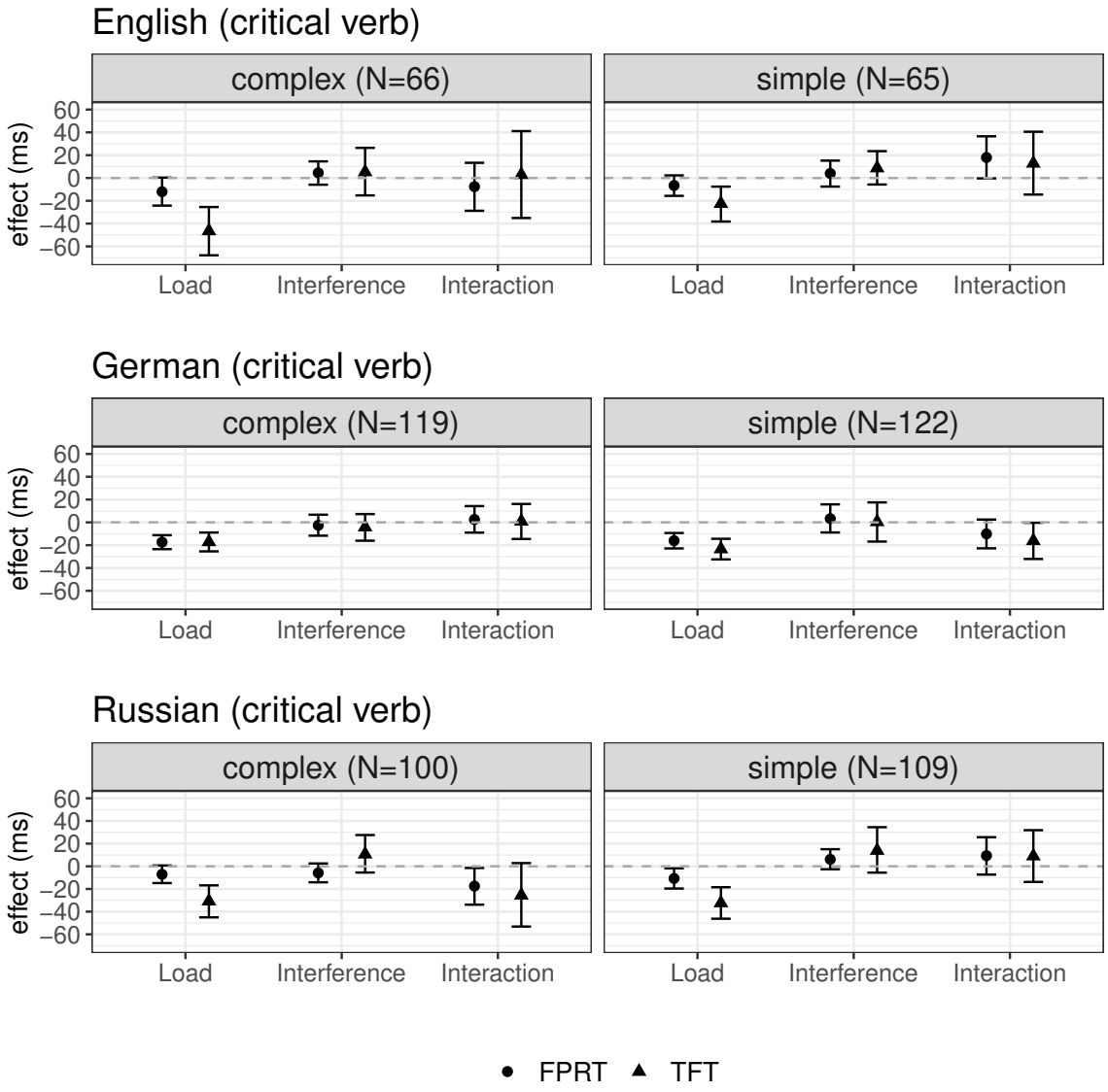

Figure 6. Effect of Load, Interference and their interaction at the critical relative clause verb for the complex and the simple versions of the English, German, and Russian experiments. Values were back-transformed from the log scale to the millisecond scale. FPRT = first-pass reading times, TFT = total fixation times.

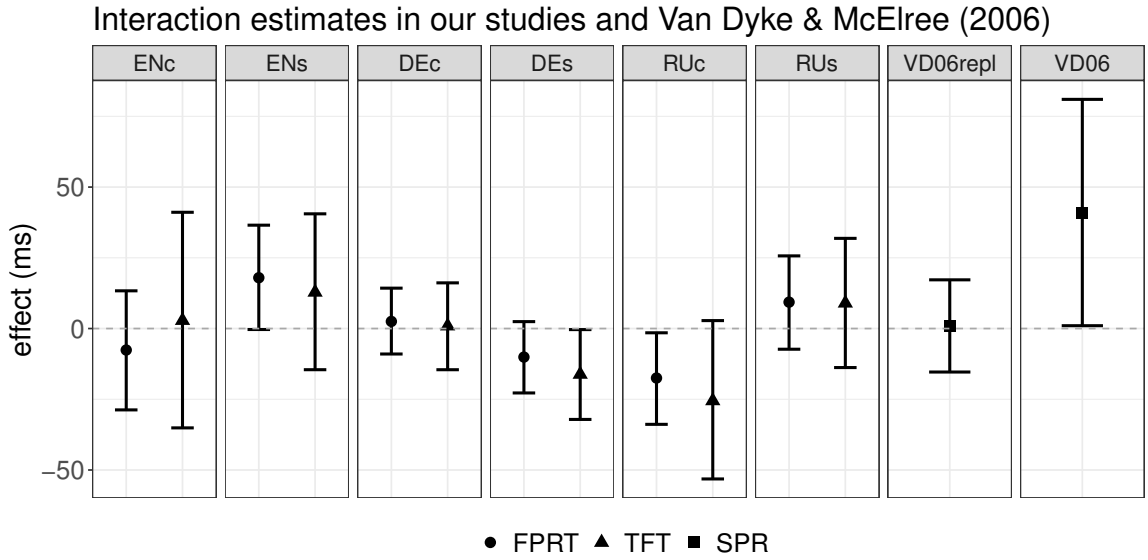

Figure 7. Shown are the interaction estimates with 95% credible intervals from our English, German and Russian eye-tracking experiments in first pass reading times (FPRT) and total reading times (TFT) (EN = English, DE = German, RU = Russian, (c) = complex version, (s) = simple version), as well as the interaction estimate from our English online self-paced reading replication ("VD06repl") of Van Dyke & McElree (2006) and the interaction estimate with 95% confidence interval from the original Van Dyke & McElree (2006) (self-paced reading times) study.

conclusions, it is important to rule out task-related explanations for these results.

Some potential concerns. In the dual-task design, the interference effects are entirely contingent on the encoding of the distractor nouns in memory. If the memory nouns are not encoded in memory, then no interference effects would be expected. This is a possibility if participants only paid close attention to the reading task but not the recall task. Although they were instructed to attend to both, if participants did not perform both tasks concurrently in our experiments, then the lack of an interaction would not necessarily bear on the theoretical question under investigation.

Our study had a lower recall accuracy, compared with the previous study by Van Dyke and McElree (2006). The low memory recall accuracy could have contributed to our study not observing support for proactive interference effects. If so, can interference effects be observed when all three memory nouns are encoded in memory? We checked this hypothesis by analyzing the English, German and Russian data with perfect recall. The 'high recall' data show similar estimates to our original analysis (see Appendix B). These results indicate that even when all memory nouns are encoded in memory, there does not seem to be much support for the hypothesis that sentence-external items interfere with within-sentence dependency resolution.

A related concern is that the dual-task paradigm could also have failed if participants paid attention only to the recall task but not to the reading task. In this

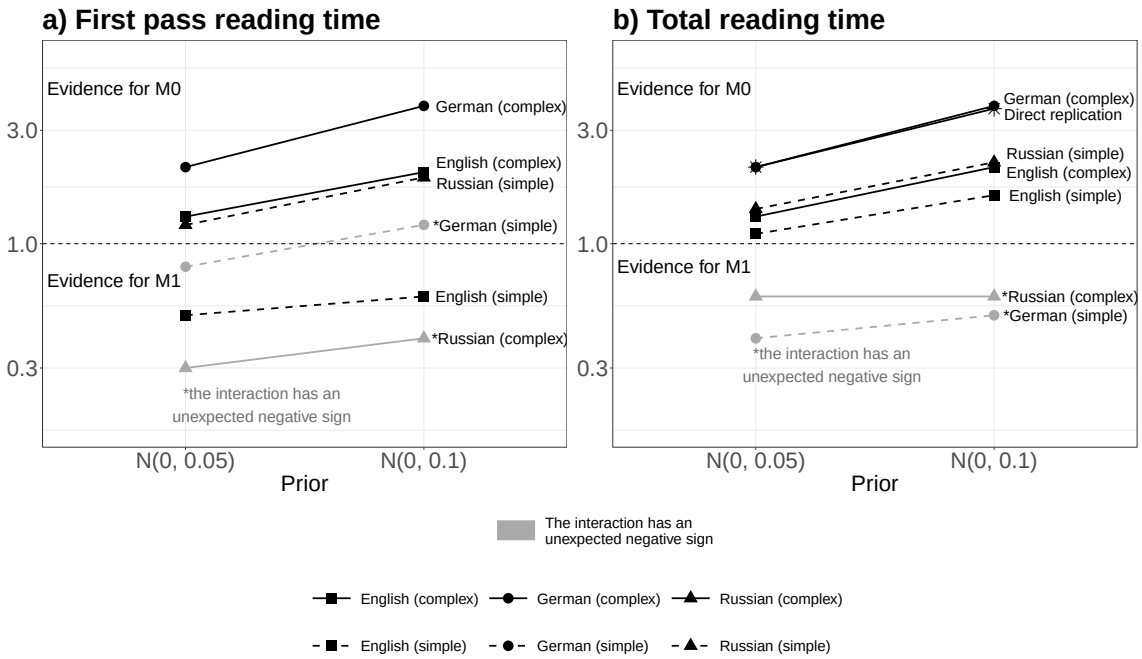

Figure 8. Bayes factor results for Model 0 over Model 1 (BF01) in a) first pass reading time and b) total reading times for the complex and the simple versions of the English, German and Russian experiments. Bayes factor values are shown for increasingly uninformative prior distributions: $\mathcal{N}(0, 0.05)$, and $\mathcal{N}(0, 0.1)$. The English SPR study results refer to our direct replication attempt of the original Van Dyke and McElree study–this study was conducted after the first three eye-tracking experiments, and is reported as Experiment 4 below.

case, we also expect to see no interference. However, the high comprehension question accuracies across all experiment versions suggest that participants attended to the reading task. In trials where recall was perfect, comprehension question accuracy is also very high (Appendix B). Moreover, in perfect recall trials, the by-condition reading times are not unusually fast which may have been an indication that the reading task was not attended to (Appendix B). This arguably minimizes the concern that the failure to observe clear proactive interference effects resulted from participants disregarding one of the two tasks in the dual-task paradigm.

A further potential concern may be that our specific stimuli may have contributed to most experiment versions not showing the predicted reading time pattern. Our study used filler-gap dependencies with two embedded relative clauses. The expectation was that these sentences may increase our chances to detect an effect compared with the object clefts in Van Dyke and McElree (2006). In the latter, the retrieval target was in linguistic focus, whereas in our sentences, it was not. Our reasoning for changing the stimuli was that items in focus have more distinctive memory representations which may reduce the magnitude of the interference effect. Conversely, it could be argued

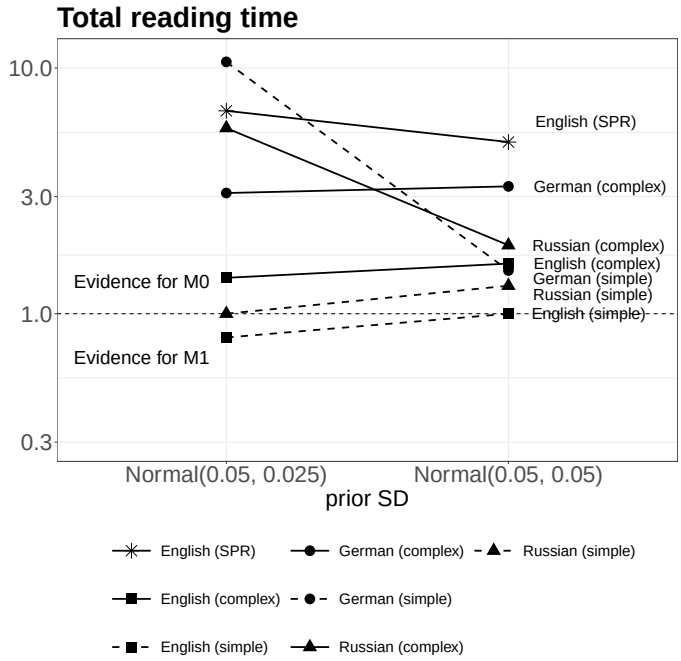

Figure 9. Bayes factor results for Model 0 over Model 1 (BF01) in total reading times for the complex and the simple versions of the English, German and Russian experiments, and Bayes factor for Experiment 4 (online SPR replication of Van Dyke and McElree, 2006). Bayes factor values are shown for enthusiastic prior distributions centered around 0.05: $\mathcal{N}(0.05, 0.025)$, and $\mathcal{N}(0.05, 0.05)$

that a target in linguistic focus, as in the clefted constructions, is less distinguishable from the prominent sentence-external distractors. In Van Dyke and McElree (2006), both the target as well as the memory nouns were given special status, which could make them more confusable. This could be the case if prominence is used as a retrieval cue (e.g., Kush et al., 2019). However, our Russian data speak against the hypothesis that clear interference effects were not observed due to the specific stimuli used here. The Russian stimuli in our study use demonstrative pronouns for the target NP, as there are no articles in Russian. This increases the prominence of the target compared to the default option (use of the NP without a demonstrative pronoun). Despite the special status of the target in Russian, there was no support for the predicted interference effect.

Nevertheless, because of the differences between our experiments and the original Van Dyke and McElree (2006) study, we additionally attempted to replicate the original Van Dyke and McElree (2006) experiment. This direct replication attempt is reported next.

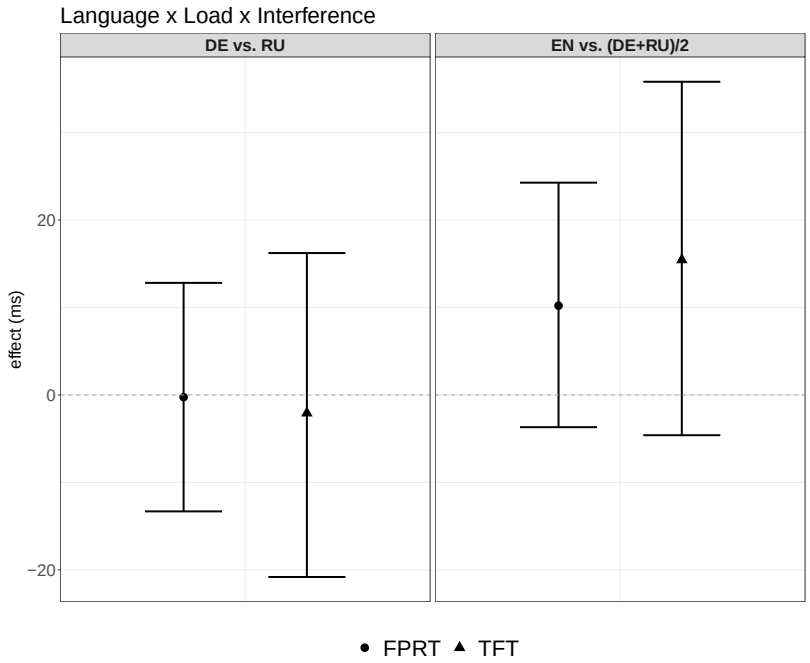

Figure 10. Shown are the posterior means and 95% credible intervals for the Language × Load × Interference interaction (first-pass reading times, total reading times). For the analysis, the combined data from all experiment versions had 304 subjects and 240 items. We tested whether the Load × Interference interaction differs in German vs. Russian, and whether it differs in English vs. German and Russian. The contrasts for the factor Language were specified using the R package `hypr` (Rabe et al., 2020), testing the null hypotheses H0_1: DE − RU = 0, i.e. whether the difference between the Load × Interference interaction in German (DE) vs. the interaction in Russian (RU) is equal to zero (German was coded as +0.5, Russian as −0.5); H0_2: EN − (DE + RU)/2 = 0, i.e., whether the difference between the Load × Interference interaction in English (EN) vs. the interaction in German and Russian (averaged) is equal to zero. English was coded as +0.66, and German and Russian as −0.33.

Experiment 4: Replication attempt of the original Van Dyke and McElree (2006) study

Our replication attempt of the self-paced reading study (SPR) by Van Dyke and McElree (2006) was implemented as an online SPR study which was implemented in PC Ibex (https://farm.pcibex.net/). The experiment was carried out online because our lab is located in Germany and finding native speakers of English locally would have been difficult.

Experimental design and materials

Our experiment used the same 2 × 2 repeated measures design that was described above: we manipulated the factors Memory load (*Load, No load*) and Interference (*No interference, Interference*) (Van Dyke & McElree, 2006).

The SPR replication attempt tested the same 36 experimental sentences as Van Dyke and McElree (2006), that is, object-cleft sentences that have the target object in linguistic focus. The sentences had the same structure as the example shown in Table 1 in the introduction. We made minor modifications to some of the original sentences to adapt them for a British English context.

Each sentence was followed by yes-or-no comprehension questions (e.g., *Did the guy live by the sea?*) with a 50:50 yes-to-no ratio. We also re-used the filler items of the original study. However, we only used a subset, 80 of the original 144 fillers, to reduce the length of the experimental session and avoid fatigue effects from this complex dual task. As in the original study, all fillers were followed by a yes-or-no comprehension question, but only half of the fillers were preceded by three memory words. All materials are available in the supplementary materials of this paper.

Participants

Participants were recruited via Prolific (https://www.prolific.co/). We tested 220 monolingual native speakers of English (from the UK and Ireland) from the age of 18 who received 8.50 GBP for their participation. During the pre-screening, participants with known speech or language related disorders were excluded. Before the experiment started, participants were asked five questions about the instructions they read. One participant was excluded because only three out of five questions were answered correctly. In addition, the data of 7 participants was lost due to some technical issues. Thus, we were able to conduct the statistical analysis with the data from 212 participants.

Procedure

Participants read the experiment instructions and answered five questions about the instructions. After reading the instructions, they were reminded that it is important to pay attention to both the self-paced reading and the recall task. Participants were shown six practice trials before the start of the experiment to familiarize them with the procedure.

Similar to the original Van Dyke and McElree (2006) study, in half of all trials participants read aloud and memorized three nouns that appeared on the screen for three seconds. They then read an experimental sentence region by region, uncovering each region by pressing the space bar on the keyboard. The sentence regions were displayed in the same manner as in the original study (It was the boat / that the guy / who / lived / by the sea / sailed / in two sunny days.). In each trial, participants were subsequently required to answer a yes-or-no comprehension question about the

sentence. Finally, for trials with memory items, these three memory words needed to be recalled (by typing them in on their keyboard). Experimental and filler sentences were shown in a randomized order. The entire experiment session took approximately 45 minutes.

Predictions and statistical analysis

As in the previous experiments, a Load × Interference interaction was expected. Like the Bayes factor analyses, the statistical analysis of the SPR data was carried out by fitting Bayesian hierarchical models using the R package `brms`.

Results of Experiment 4

Comprehension question accuracy. Figure 11 shows the comprehension question accuracies by condition. The comprehension question accuracies were overall high and in the same range as in our English eye-tracking experiments (see Figure 11) and the Van Dyke and McElree (2006) response accuracies.

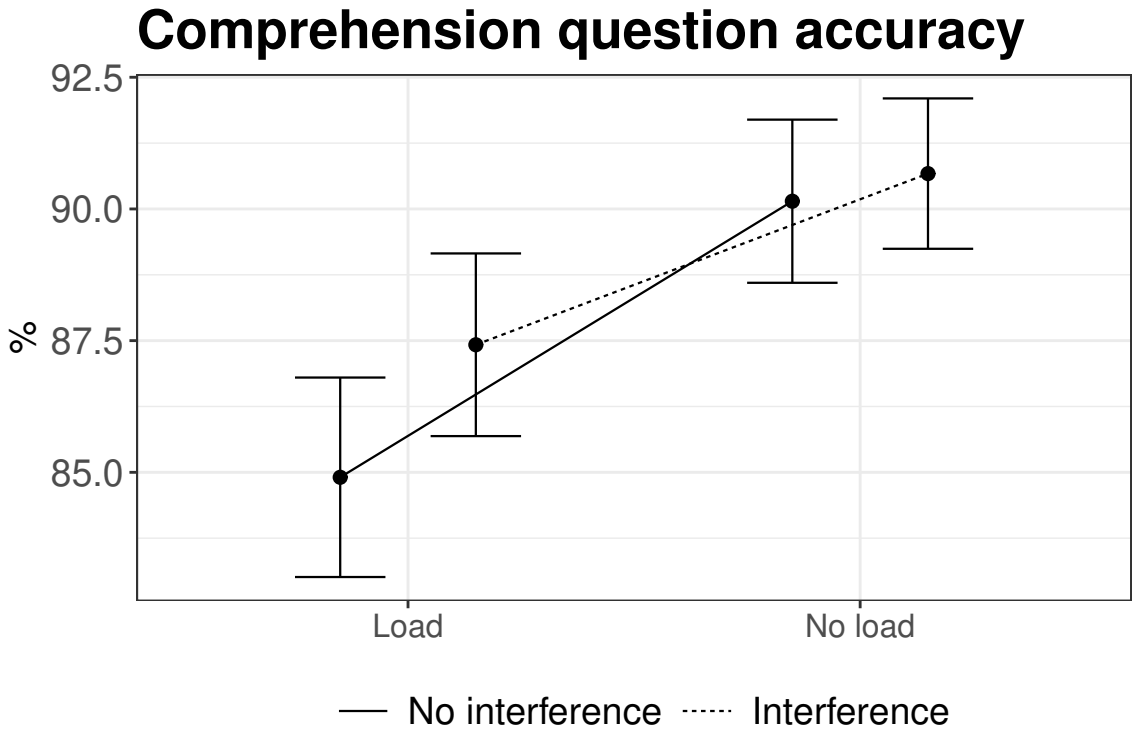

Figure 11. By-condition mean comprehension question accuracies (in percent) with 95% confidence intervals for Experiment 4, the attempted replication of Van Dyke & McElree (2006).

Memory recall accuracy. Figure 12 shows the memory recall accuracies for the SPR replication attempt. For a strict accuracy criterion (that is, all three words had to be recalled in the correct order), our Experiment 4 shows the following 95% confidence intervals (CIs): For No Interference conditions the interval is [69, 76]%, for Interference conditions, it is [68, 75]%. These recall accuracies are somewhat lower than the recall accuracies in the original Van Dyke study (No Interference [75, 83]%, Interference [73, 81]%).

In the section Some potential concerns, we discussed potential issues that could have arisen in our eye-tracking data which may have prevented us from observing the expected interaction: if participants paid attention to the reading task but not the memory recall task, or vice versa, no interference would be expected.

As was shown in Figure 12, the memory recall was again somewhat lower in our Experiment 4 compared to the original Van Dyke and McElree (2006) study. A subset of participants had extremely low recall accuracies. Under the strict criterion, three participants even had an accuracy of 0% in the recall task for experimental items. We therefore carried out two sets of analyses, one with all the participants included, and another analysis with only those participants who showed a very high recall accuracy under the strict criterion as well as a very high comprehension accuracy(both >80%).

The first analysis, which included all participants, is more similar to the analysis that Van Dyke and McElree (2006) reports because the authors did not exclude any participants who had low accuracy. Our second analysis, with only those participants who had high recall accuracy, guarantees that participants were holding all three sentence-external items for a given trial in memory; this ensures that any interference effect has a higher chance of being detected. The additional high comprehension accuracy ensured that these participants also paid attention to the reading task.

Self-paced reading times: Analysis with all participants included. Figure 13 shows the by-condition means with 95% confidence intervals at the critical verb region of our Experiment 4 (left panel) and the original Van Dyke & McElree (2006) study (right panel).

In Figure 14, we show the effects of Load, Interference and their interaction at the critical region. The estimate of interest, the interaction, is centered on zero (95% CrI [−15, 17] ms). Here, we refer the reader back to Figure 7 which shows the interaction of interest at the critical verb region next to the interaction estimates from the original Van Dyke & McElree study and our English, German and Russian eye-tracking experiments.

Table 7 suggests anecdotal to moderate evidence in favor of Model 0 over Model 1. That is, the direct replication Experiment 4 furnishes evidence *against* the Load × Interference interaction. Figures 8 and 9 show the Bayes factor results of Experiment 4 (the direct replication attempt) alongside the Bayes factor results of our eye-tracking experiments.

Self-paced reading times: Analysis of participants with a high recall accuracy. As mentioned in the Memory recall accuracy section, we conducted an

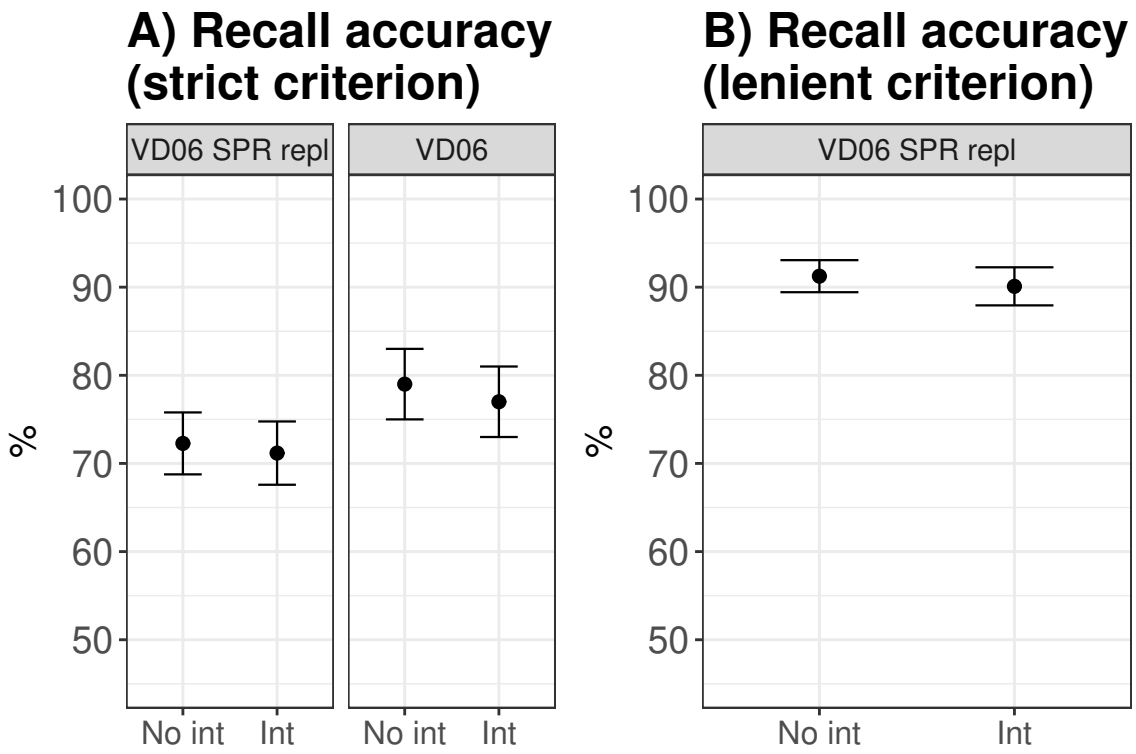

Figure 12. Mean recall accuracies (in percent) with 95% confidence intervals, A) for a strict criterion (all three words recalled in the correct order) from the online self-paced reading replication attempt of Van Dyke & McElree (2006) and the original Van Dyke & McElree (2006) experiment, and B) for a lenient criterion (two or three words recalled in any order) for our replication attempt. No int = No interference condition, Int = Interference condition.

additional analysis of the data from a subset of participants (n = 90) that had both a question response accuracy of $\geq$ 80% (95% CIs for conditions Load/Interference [88, 93]%, Load/No interference [85, 91]%, No Load/Interference [91, 95]%, No Load/No Interference [91, 95]%) and a recall accuracy of $\geq$ 80% (95% CIs for conditions Load/Interference [91, 94]%, Load/No interference [91, 95]%).

The Load $\times$ Interference interaction estimate of the reading time analysis had a 95% CrI ranging from -31 to 26 ms. Thus, our online self-paced reading replication of Van Dyke and McElree (2006) shows no indication of the Load $\times$ Interference interaction, even when participants seemingly paid close attention to both the reading and the recall task.

General Discussion

The main goal of this study was to investigate cue-based retrieval theories' prediction of cue-based retrieval interference from sentence-external distractors. The

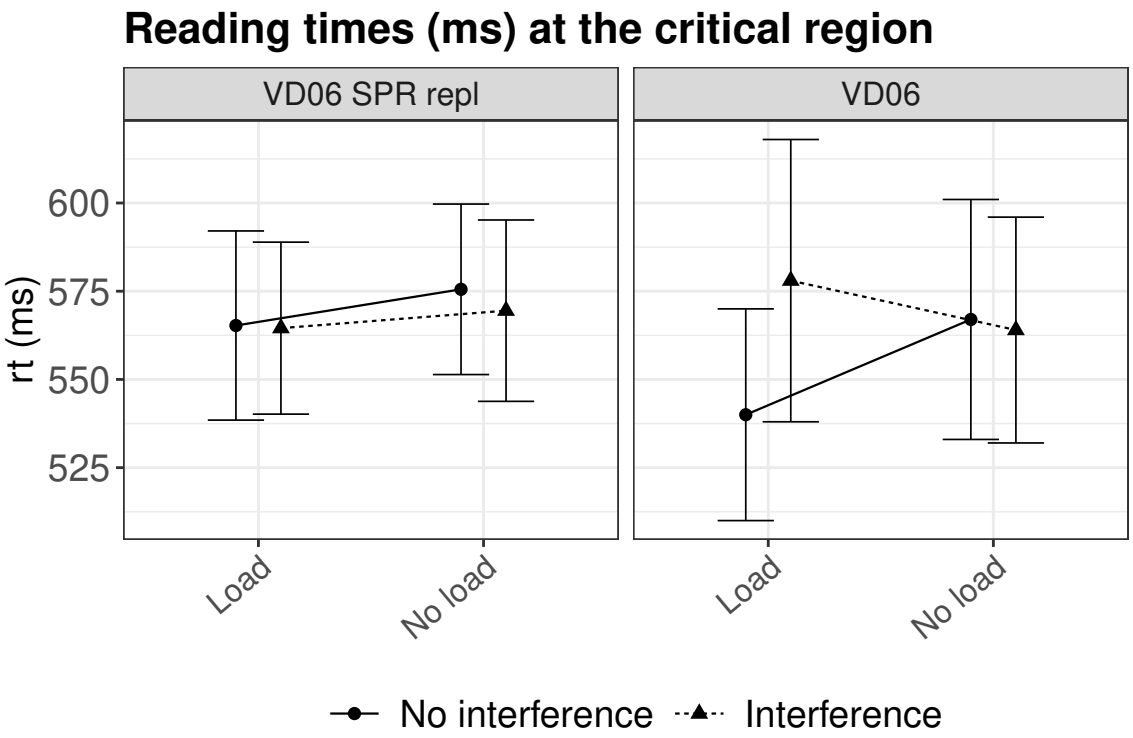

Figure 13. By-condition mean reading times with 95% confidence intervals at the critical region for Experiment 4, the attempted replication of Van Dyke & McElree (2006) as well as for the original Van Dyke & McElree (2006) experiment.

present study was the first to test proactive similarity-based retrieval interference cross-linguistically, using eye-tracking as well as self-paced reading, and with an experiment design that follows that of Van Dyke and McElree (2006). We aimed to extend previous findings on English object clefts to the online processing of filler-gap dependencies (Van Dyke & McElree, 2006). Our relatively large sample sizes in German and Russian and the English SPR study provided more precise estimates of the critical Load × Interference interaction that has been the focus of previous work in this literature. A further novel contribution of this study is a within-subjects manipulation of processing depth, investigating proactive interference under varying task demands.

By and large, we found that there was compelling evidence against the critical interaction of Load × Interference: This was clearly the case in German and Russian, as well as in our replication of Van Dyke and McElree (2006). This lack of an interaction effect suggests that the presence of semantically-associated distractors in an extra-sentential list did not routinely impact online dependency formation. However, our analyses did reveal some weak evidence in favor of the predicted Load × Interference interaction in the simple version of the English experiment. Thus, only the English eye-tracking data lend any support to the hypothesis that semantically

Table 7
Bayes factor results for Model 0 over Model 1 (BF01) for Experiment 4, the direct replication attempt of the original Van Dyke and McElree (2006) study. Bayes factor values are shown for increasingly informative prior distributions for the models of the statistical analysis that includes the data of all participants and the analysis that includes the data of participants with high recall accuracy.

	Bayes factor (evidence for M0)	Posterior estimate of the interaction (log-ms)
All participants (N = 212)		
Enthusiastic priors		
N(0.05, 0.025)	6.7	0.02 [-0.01, 0.04]
N(0.05, 0.05)	5	0.01 [-0.02, 0.04]
Mildly informative priors		
N(0, 0.05)	3.3	0 [-0.03, 0.03]
N(0, 0.1)	6.1	0 [-0.03, 0.03]
Participants with high recall accuracy (N = 90)		
Enthusiastic priors		
N(0.05, 0.025)	4.0	0.02 [-0.01, 0.06]
N(0.05, 0.05)	3.4	0.01 [-0.04, 0.05]
Mildly informative priors		
N(0, 0.05)	2.1	-0.001 [-0.05, 0.04]
N(0, 0.1)	3.7	-0.001 [-0.05, 0.05]

similar, sentence-external distractors in memory can interfere with retrieval during real-time sentence comprehension, and even then, only in testing conditions that promote shallow processing.

However, our direct self-paced reading replication attempt of the original study (our Experiment 4) showed moderate to strong evidence against the critical interaction, again suggesting that proactive interference effects do not generally impact online dependency formation in English in all testing contexts. Finally, it is notable that German and Russian did not show evidence for this proactive retrieval interference effect, despite using the same method and design, and comparable sentence structures across languages.

Taken together, the four experiments reported provide some cross-linguistic evidence against the claim that proactive retrieval interference from sentence-external material can interfere with within-sentence dependency resolution.

Is proactive interference modulated by language or task demand?

It is unclear why the evidence for proactive interference was limited only to English speakers under low task demands. One explanation is that this result is simply due to noise in the data (i.e., a Type I error). Given that—of all experiment versions—only

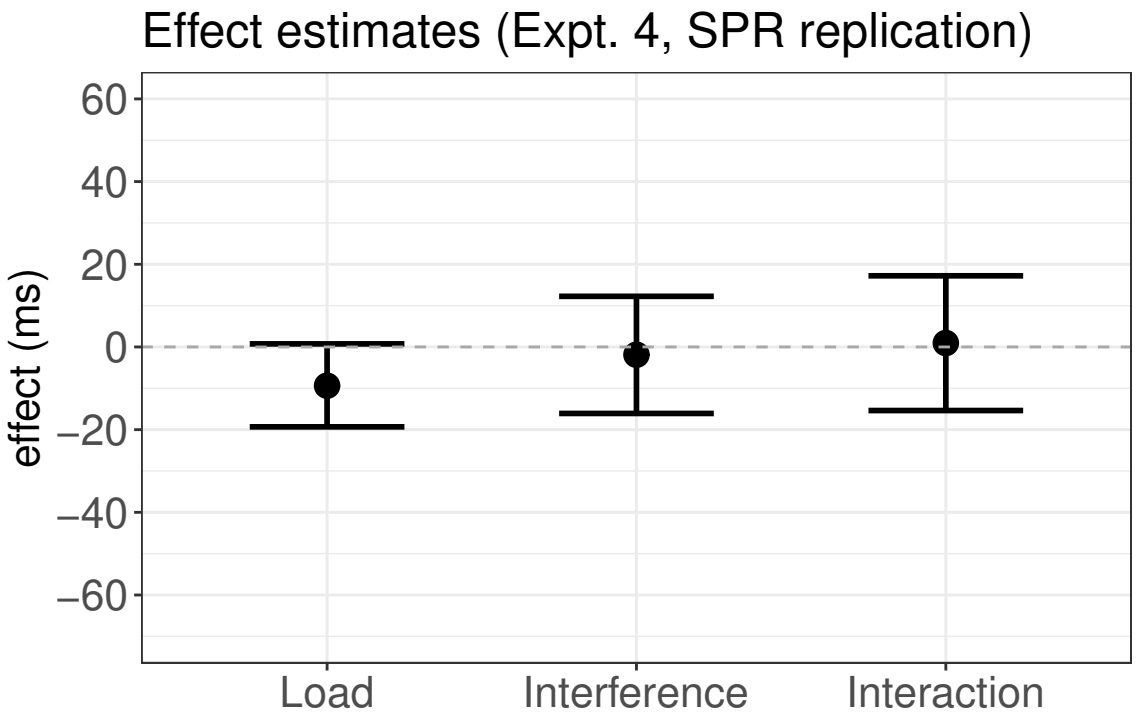

Figure 14. Effect of Load, Interference and their interaction at the critical relative clause verb for the online self-paced reading replication attempt of Van Dyke & McElree (2006). Values were back-transformed from the log scale to the millisecond scale.

the lower-sample size English study (simple version, FPRT) showed any evidence for proactive interference, this could be an accidental outcome that may not be replicable in a future study. Before too strongly interpreting the theoretical implications of this finding, it would be important to show that proactive retrieval interference is replicable under a low processing load in a large-sample confirmatory experiment (e.g., de Groot, 2014; Nicenboim et al., 2018; Vasishth et al., 2018).

If the effect is not due to noise in the data, the weak evidence for interference in English and evidence against interference in German and Russian may point to cross-linguistic processing differences. The Language × Load × Interference interaction is inconclusive and at best weakly supports this; the 95 % credible interval largely has a positive sign but also includes negative values. In German and Russian, the accusative case marking of the relative pronoun/complementizer referring to the grammatical object may make the retrieval target more distinguishable, and therefore eliminate interference during online dependency formation. Whether this is indeed the case is unclear at present and requires further investigation using a cross-linguistic design that explicitly manipulates the diagnosticity or presence of case marking.

However, explanations that appeal to differences in how diagnostic language-specific

retrieval cues do not address the question of why this occurred only in low demand testing conditions in English. Our depth of processing hypothesis postulated that inducing shallow processing would lead to less or no interference during online sentence comprehension (Logačev & Vasishth, 2016a; Swets et al., 2008). However, this was opposite to the pattern we saw in our data, with only the simple version showing some support for proactive interference. In the complex version, the effect was not replicated in the sense that we did not see observing overlapping effect estimates with the predicted positive sign (the region of practice equivalence approach, Kruschke & Liddell, 2018).

One explanation for this pattern might be that this effect only arises under good-enough processing conditions. Laurinavichyute and von der Malsburg (2021) recently found support for agreement attraction effects under superficial but not deep processing conditions. The patterns observed in our study (English, simple) would be consistent with the suggestion that only superficial processing leads to interference effects.

It is also possible that in our study, the difference in processing depth does not reflect superficial vs. deep processing as we had hypothesized, but rather "typical" vs. deep processing. That is, the simple questions conditions may not have resulted in underspecification of syntactic dependencies, as intended. Rather, in this version, participants may have engaged in "typical," attentive sentence processing. The formal lab setting and the high comprehension question accuracies would speak in favor of this possibility. By comparison, the more complex questions may have induced a more unnatural, heightened attention state for the participants. Under this view, interference may occur in "typical" language processing mode, in which the parser attempts to resolve all dependencies. However, interference does not arise in a deep processing mode when particular attention is required, such as maintaining each dependency in memory to correctly answer a comprehension question.

Broader implications of the current study

Setting aside the question of why English low-demand testing conditions present an apparent exception, our broader conclusion that proactive interference from sentence-external distractors is not observed in many testing conditions holds important conclusions for cue-based retrieval theory. Why don't we observe the predicted reading time pattern, with possibly one exception (the English eye-tracking data)?

One possibility is that the relevant effect has too small an effect on processing to be observed even with the sample sizes collected here. As foreshadowed above, interference in sentence processing is likely to be a subtle phenomenon that is difficult to detect. Recall that the meta-analysis by Jäger et al. (2017) on non-agreement subject-verb dependencies showed an interference effect estimate of 13 ms with a 95% credible interval ranging from 2 to 28 ms. Similarly, a self-paced reading study on number interference by Nicenboim et al. (2018) with a participant sample size of 182 showed an interference effect estimate of 9 ms (95% CrI [0,18] ms). Our English experiment obtained a similarly precise, and small, estimate of the effect of interest in

first-pass reading times (95% CrI [0,18] ms). However, with the rather small sample size, we are in a relatively low-power scenario. Even with larger participant samples, as in German or Russian, such small effects may remain undetected. Consequently, detecting proactive interference effects would require significantly more resources, or more sensitive methods.

This study may also show little support for proactive interference because proactive interference effects are overall weaker than retroactive interference effects. Van Dyke and McElree (2011) established, through a within-subjects/items design, that proactive interference (from within-sentence distractors) is overall weaker than retroactive interference. This finding is supported by a meta-analysis result in Jäger et al. (2017), showing that in general, proactive interference leads to a smaller reading time slowdown than retroactive interference.

Proactive interference from sentence-*external* materials could be even more subtle. Distractors may predominantly play a role when they appear within a sentence (e.g., Mertzen et al., 2023; Rich & Wagers, 2020; Van Dyke, 2007), particularly when the distractor intervenes linearly between the retrieval target and retrieval site. This would be expected if, for example, the retrieval cues used to reactivate the filler noun phrase strongly distinguish within-sentence material from other material in working memory. This theoretical possibility has been suggested by authors who proposed that abstract clause-bounding retrieval cues are used to guide retrieval (Wagers et al., 2009). There is also some empirical evidence that is compatible with this idea: Dillon et al. (2017) showed that retrieval interference is diminished when the distractors are inside appositive relative clauses, which may be seen as a parenthetical distinct from the target sentence. These studies suggest that the context of encoding of the distractor—inside a restrictive relative clause, or inside a more syntactically independent appositive relative clause—may modulate the degree of retrieval interference observed. If the context of encoding a distractor element is critical, then cue-based retrieval theories might predict only very small, if any, interference from sentence-external distractors of the sort tested here.

Conclusions

Our data suggest that extra-sentential items encoded in memory do not interfere with within-sentence dependency formation. Taken together, our data from four experiments present cross-linguistic evidence against proactive interference from extra-sentential items. The broader implication of this finding is that similarity-based interference may be limited to linguistic items that are part of the linguistic context in which they appear; distractors must be linguistically relevant in order to cause interference.

Acknowledgements

We are grateful to our lab manager Johanna Thieke (University of Potsdam, Germany) and the following (former) research assistants for their help with data collection for the German experiment: Alexandra Lorson, Maria Korochkina, Elna Haffner, Marie de la Fuente, Romy Leue, Luzie Ahrens, Hanna Eversmann, Chiara Tschirner. For the English experiment (UMass Amherst, USA): Austin Tero and for the Russian experiment (Higher School of Economics Moscow, Russia): Maria Ignashina. Thanks also go to Hiroki Fujita for helpful comments.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 317633480 – SFB 1287, Project B03.

Author contributions

- Daniela Mertzen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project administration
- Anna Laurinavichyute: Software, Formal analysis, Data Curation, Writing – Review & Editing
- Brian W. Dillon: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Project administration
- Ralf Engbert: Conceptualization, Funding acquisition, Supervision, Writing – Review
- Shravan Vasishth: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing – Review & Editing, Supervision, Funding acquisition

Appendix A
Experimental materials (German & Russian)

German stimuli

We created 40 experimental items for the complex and 40 items for the simple version of the experiment. Table A1 shows two example items, one followed by a complex and one followed by a simple comprehension question. The critical dependency is between the relative clause verb *steuerte/reparierte* ('steered'/'repaired') and the sentence-initial NP *Das Boot* ('The boat').

The German stimuli differed from the English materials in the following way: In Figure 2, we showed that for German, the relative pronoun is overtly marked for accusative case whereas in English, the complementizer *that* is not. This is an important distinction, because in German, case could serve as an additional cue for retrieval and, hence, reduce interference. For half of the items, the sentence-initial target noun phrase is of masculine grammatical gender such that the relative pronoun of the following object relative clause is unambiguously marked for accusative case (Figure 2). Note, however, that for the other half of the items, the target noun phrase is neuter such that the surface form of the relative pronoun is ambiguous between nominative and accusative case, as in the example in Table A1. To increase naturalness of the German sentences, an adverb preceded the critical relative clause verb because a prepositional- or adverbial phrase, as was added in English, is not licensed in the post-verbal position in German.

Russian stimuli

The Russian study followed the same design as the English and German study. The complex and simple version each had 40 items. Table A2 shows two example items. The critical dependency is between the relative clause verb сломала/присмотрела ('broke'/'found') and the sentence-initial NP То кресло (The armchair).

Differences between Russian vs. English and German are the following: An adverb was added after the critical relative clause verb for greater naturalness of the sentence. As there are no articles in Russian, the demonstrative pronouns тот (masculine), та (feminine) and то (neuter) (*that*) were used for the sentence-initial NP. The demonstratives make the NP more prominent compared to the default option of the NP without a demonstrative pronoun, and in comparison with our English and German study. This is similar to the Van Dyke and McElree (2006) study which used object cleft sentences that increase the prominence of the target NP. Case marking on the relative pronouns is overt in Russian, although for masculine and neuter the surface form is case ambiguous. One third of items had a feminine, one third a masculine, and another third a neuter target NP.

Table A1
German example items (complex and simple version).

A) Memory load:	Kühlschrank *fridge*	Waschmaschine *washing machine*	Computer *computer*

No interference

Das Boot, das der Mann, der am Meer lebte, gestern *steuerte*, schien schon alt zu sein.
The boat, that the man, who at sea lived, yesterday steered, seemed already old to be.

Interference

Das Boot, das der Mann, der am Meer lebte, gestern reparierte, schien schon alt zu sein.
The boat, that the man, who at sea lived, yesterday repaired, seemed already old to be.

No memory load: — — —

No interference

Das Boot, das der Mann, der am Meer lebte, gestern *steuerte*, schien schon alt zu sein.
The boat, that the man, who at sea lived, yesterday steered, seemed already old to be.

Interference

Das Boot, das der Mann, der am Meer lebte, gestern *reparierte*, schien schon alt zu sein.
The boat, that the man, who at sea lived, yesterday repaired, seemed already old to be.

'The boat that the man who lived by the sea steered/repaired seemed to be quite old.'
'Complex' question: *'Hat der Mann am Meer gelebt?' (Did the man live by the sea'?)*

B) Memory load:	Parfüm *perfume*	Rauch *smoke*	Leder *leather*

No interference

Der Kaffee, den der Genießer, der in der Rösterei saß, gerne *trank*, schien äußerst aromatisch zu sein.
The coffee, that the connoisseur, who in the roastery sat, gladly drank, seemed most aromatic to be.

Interference

Der Kaffee, den der Genießer, der in der Rösterei saß, gerne *roch*, schien äußerst aromatisch zu sein.
The coffee, that the connoisseur, who in the roastery sat, gladly smelled, seemed most aromatic to be.

No memory load: — — —

No interference

Der Kaffee, den der Genießer, der in der Rösterei saß, gerne *trank*, schien äußerst aromatisch zu sein.
The coffee, that the connoisseur, who in the roastery sat, gladly drank, seemed most aromatic to be.

Interference

Der Kaffee, den der Genießer, der in der Rösterei saß, gerne *roch*, schien äußerst aromatisch zu sein.
The coffee, that the connoisseur, who in the roastery sat, gladly smelled, seemed most aromatic to be.

'The coffee that the connoisseur who sat in the roastery drank/smelled seemed to be most aromatic.' 'Simple' question: *'Wurde in diesem Satz eine Rösterei erwähnt?' (Was a roastery mentioned in this sentence?)*

Table A2
Russian example items (complex and simple version).

A) Memory load:	ковер	подушка	покрывало
	carpet	pillow	cover
No interference			
То кресло, которое старушка, любящая антиквариат сломала недавно, относится к . . .			
The armchair, that elderly woman, loving antiques broke recently, belongs to . . .			
Interference			
То кресло, которое старушка, любящая антиквариат присмотрела недавно, относится к . . .			
The armchair, that elderly woman, loving antiques found recently, belongs to . . .			
No memory load:	—	—	—
No interference			
То кресло, которое старушка, любящая антиквариат сломала недавно, относится к . . .			
The armchair, that elderly woman, loving antiques broke recently, belongs to . . .			
Interference			
То кресло, которое старушка, любящая антиквариат присмотрела недавно, относится к . . .			
The armchair, that elderly woman, loving antiques found recently, belongs to . . .			

'The armchair that the elderly woman who loves antiques recently broke/found belongs to the Chippendale style.' Complex question: Did the elderly woman love antiques?

B) Memory load:	волна	камень	яхта
	wave	stone	yacht
No interference			
Тот матрас, который акула, плавающая в море, прокусила неожиданно для туриста, принадлежит. . .			
The mattress, that shark, swimming in sea bit through surprisingly for tourist belongs to. . .			
Interference			
Тот матрас, который акула, плавающая в море, заметила неожиданно для туриста, принадлежит. . .			
The mattress, that shark, swimming in sea spotted surprisingly for tourist belongs to. . .			
No memory load:	—	—	—
No load, No interference			
Тот матрас, который акула, плавающая в море, прокусила неожиданно для туриста, принадлежит. . .			
The mattress, that shark, swimming in sea bit through surprisingly for tourist belongs to. . .			
No load, Interference			
Тот матрас, который акула, плавающая в море, заметила неожиданно для туриста, принадлежит. . .			
The mattress, that shark, swimming in sea spotted surprisingly for tourist belongs to. . .			

'The inflatable mattress that the shark swimming in the sea surprisingly spotted/bit through, belonged to the tourist's daughter.
Simple question: Was an ophthalmologist mentioned in the sentence?

Filler items

English. For the English study, 50 "true" filler items of two varying syntactic structures were created for each of the two experiment versions. 25 were object cleft constructions and 25 were short, simple sentences starting with a quantifier. A further 40 filler items came from another experiment that tested the processing of object- vs. subject relative clauses. All filler items were followed by a comprehension question, and half of all fillers were preceded by three memory nouns.

German. 90 filler items were created for each of the two versions. 30 sentences were object cleft-sentences with two embedded relative clauses (the outer relative clause being an object relative clause, the most embedded relative clause being a subject relative clause). Here half of the questions targeted the object-verb dependency between the sentence-initial noun phrase and the object relative clause verb. A further 30 were subordinate clause–main clause constructions starting with the subordinate conjunction *that*, and the remaining 30 filler sentences were subject-extracted RC structures.

Russian. For each of the two versions, 90 filler items of varying syntactic structures were created. 30 sentences were object cleft-sentences with an embedded relative clause. A further 30 were subordinate clause–main clause constructions starting with a subordinate conjunction expressing causal relationships, and the remaining 30 filler sentences were sentences of various structures randomly selected from the Russian National Corpus (http://www.ruscorpora.ru/).

Appendix B
Perfect recall data

Comprehension question accuracy in perfect recall trials

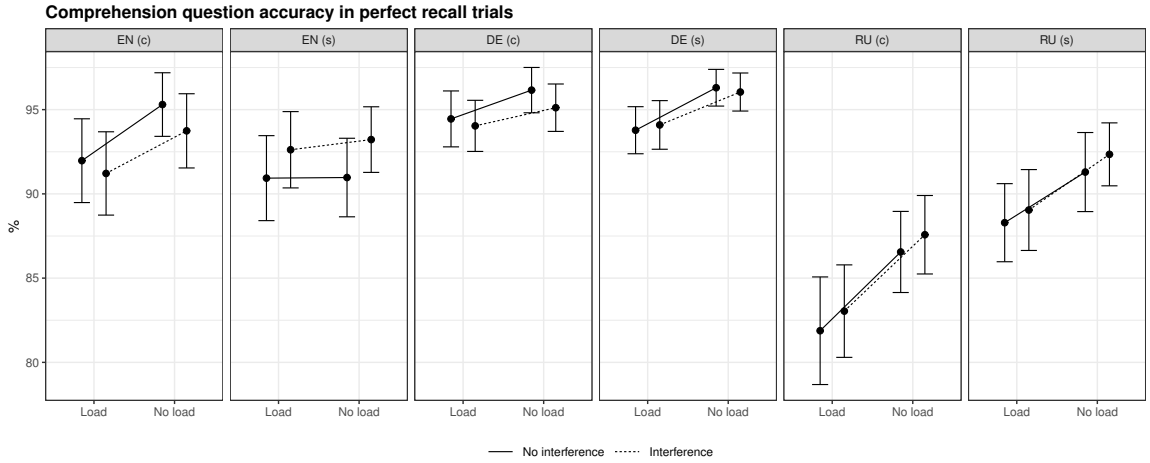

Figure B1. By-condition mean comprehension question accuracies (in percent) with 95% confidence intervals in perfect recall trials. EN = English, DE = German, RU = Russian; (c) = complex version, (s) = simple version

Raw by-condition reading times in perfect recall data

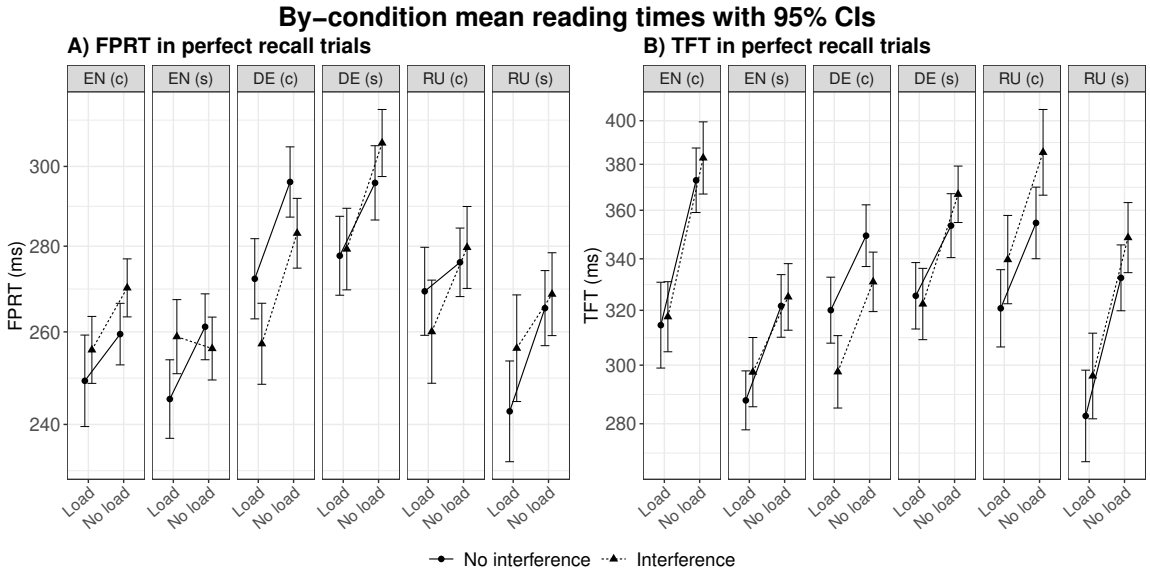

Figure B2. By-condition mean reading times with 95% confidence intervals. In (A) and (B), FPRT and TFT are shown for trials with perfect recall, i.e., recall of three memory nouns in the correct order. EN = English, DE = German, RU = Russian; (c) = complex version, (s) = simple version.

Reading time analyses for data with perfect recall

Study	Load	Interference	Interaction
	TFT Posterior mean [95% credible intervals]		
English, complex	-54 ms [-77, -30]	7 ms [-19, 33]	8 ms [-41, 54]
English, simple	-27 ms [-45, -8]	2 ms [-13, 19]	1 ms [-33, 34]
German, complex	-21 ms [-30, -12]	-4 ms [-16, 7]	1 ms [-17, 19]
German, simple	-26 ms [-37, -15]	6 ms [-13, 24]	-11 ms [-29, 8]
Russian, complex	-34 ms [-49, -19]	15 ms [-2, 32]	-18 ms [-51, 16]
Russian, simple	-42 ms [-56, -28]	9 ms [-10, 29]	0 ms [-24, 26]

Table B1
Total fixation times results for the subset of the data with three recalled memory nouns: Effect of Load, Interference and their interaction for both the complex and the simple version of the English, German and Russian experiment.

Study	Load	Interference	Interaction
	FPRT Posterior mean [95% credible intervals]		
English, complex	-12 ms [-27, 3]	4 ms [-9, 16]	-9 ms [-33, 15]
English, simple	-9 ms [-20, 1]	3 ms [-8, 15]	16 ms [-6, 39]
German, complex	-19 ms [-26, -11]	-4 ms [-13, 5]	1 ms [-13, 15]
German, simple	-17 ms [-25, -9]	8 ms [-5, 21]	-3 ms [-19, 12]
Russian, complex	-8 ms [-17, 1]	-3 ms [-12, 6]	-12 ms [-30, 5]
Russian, simple	-15 ms [-25, -5]	7 ms [-2, 17]	12 ms [-7, 30]

Table B2
-First-pass reading times results for high recall accuracy: Effects of Load, Interference and their interaction for both the complex and the simple version of the English, German and Russian experiment.

Appendix C
*

References

Arnett, N., & Wagers, M. (2017). Subject encodings and retrieval interference. *Journal of Memory and Language*, *93*, 22–54. https://doi.org/10.1016/j.jml.2016.07.005

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

Bhatia, S., & Dillon, B. W. (2022). Processing agreement in Hindi: When agreement feeds attraction. *Journal of Memory and Language*, *125*, 104322. https://doi.org/10.1016/j.jml.2022.104322

Birch, S., & Rayner, K. (1997). Linguistic focus affects eye movements during reading. *Memory & Cognition*, *25*, 653–660. https://doi.org/10.3758/bf03211306

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.09.015

Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*, 1–37. https://doi.org/10.18637/jss.v076.i01

Chomsky, N. (1971). Deep structure, surface structure and semantic interpretation. In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (pp. 183–216). Cambridge, Cambridge University Press.

Chromý, J., Brand, J. L., Laurinavichyute, A., & Lacina, R. (2023). Number agreement attraction in Czech and English comprehension: A direct experimental comparison. *Glossa Psycholinguistics*, *2*(1). https://doi.org/10.5070/G6011235

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

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

de Groot, A. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh and Han L.J. van der Maas]. *Acta Psychologica*, *148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001

Dillon, B. W., Chow, W.-Y., & Xiang, M. (2016). The relationship between anaphor features and antecedent retrieval: Comparing Mandarin ziji and ta-ziji. *Frontiers in psychology*, *6*, 1966. https://doi.org/10.3389/fpsyg.2015.01966

Dillon, B. W., Clifton, C., Sloggett, S., & Frazier, L. (2017). Appositives and their aftermath: Interference depends on at-issue vs. not-at-issue status. *Journal of Memory and Language*, *96*, 93–109. https://doi.org/10.1016/j.jml.2017.04.008

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. https://doi.org/10.1016/j.jml.2013.04.003

Engelmann, F., Jäger, L. A., & Vasishth, S. (2020). The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science*, *43*, e12800. https://doi.org/10.1111/cogs.12800

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain specific resources. *Journal of Memory and Language*, *54*, 541–553. https://doi.org/10.1016/j.jml.2005.12.006

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

Fodor, J. D., Nickels, S., & Schott, E. (2016). Center-embedded sentences: What's pronounceable is comprehensible. In R. de Almeida & L. Gleitman (Eds.), *On concepts, modules, and language: Cognitive science at its core*. Oxford University Press.

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F', and min F'. *Journal of Verbal Learning and Verbal Behavior*, *15*(2), 135–142. https://doi.org/10.1016/0022-5371(76)90014-1

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd). Boca Raton, FL, Chapman; Hall/CRC.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *6*(9), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555

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.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science*, *13*(5), 425–430. https://doi.org/10.1111/1467-9280.00475

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). Bridgesampling: An r package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Hartsuiker, R. J., Schriefers, H. J., Bock, K., & Kikstra, G. M. (2003). Morphophonological influences on the construction of subject-verb agreement. *Memory and Cognition*, *31*, 1316–1326. https://doi.org/10.3758/BF03195814

Husain, S. et al. (2021). Revisiting anti-locality effects: Evidence against prediction-based accounts. *Journal of Memory and Language*, *121*, 104280. https://doi.org/10.1016/j.jml.2021.104280

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, *6*(617). https://doi.org/10.3389/fpsyg.2015.00617

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*. https://doi.org/10.1016/j.jml.2019.104063

Jeffreys, H. (1961). *Theory of probability*. Oxford, Clarendon Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Keshev, M., & Meltzer-Asscher, A. (2019). Distant relatives: Resumptive pronouns can inherit agreement features of implied antecedents, In *32nd Annual CUNY Conference on Human Sentence Processing*.

Koesterich, N., Keshev, M., Shamai, D., & Meltzer-Asscher, A. (2021). Interference in the comprehension of filler-gap and filler-resumptive dependencies, In *34th Annual CUNY Conference on Human Sentence Processing*.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29(6)*, 627–645. https://doi.org/10.1023/A:1026528912821

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam, Academic Press.

Kruschke, J., & 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. https://doi.org/10.3758/s13423-016-1221-4

Kush, D., Johns, C. L., & Van Dyke, J. A. (2019). Prominence-sensitive pronoun resolution: New evidence from the speed-accuracy tradeoff procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(7), 1234–1251. https://doi.org/10.1037/xlm0000646

Lacina, R., & Chromý, J. (2022). No agreement attraction facilitation observed in Czech: Not even syncretism helps, In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

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

Laurinavichyute, A., Jäger, L. A., Akinina, Y., Roß, J., & Dragoy, O. (2017). Retrieval and encoding interference: Cross-linguistic evidence from anaphor processing. *Frontiers in Psychology*, *8*, 965. https://doi.org/10.3389/fpsyg.2017.00965

Laurinavichyute, A., & von der Malsburg, T. (2021). Agreement attraction in grammatical sentences arises only in the good-enough processing mode, In *34th Annual CUNY Conference on Human Sentence Processing*.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Lewandowsky, S., Geiger, S. M., & Oberauer, K. (2008). Interference-based forgetting in verbal short-term memory. *Journal of Memory and Language*, *59*(2), 200–222. https://doi.org/10.1016/j.jml.2008.04.004

Lewis, R. L. (2000). Specifying architectures for language processing: Process, control, and memory in parsing and interpretation. In M. W. Crocker, M. Pickering, & C. Clifton Jr. (Eds.), *Architectures and Mechanisms for Language Processing* (pp. 56–89). Cambridge, Cambridge University Press.

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

Logačev, P., & Vasishth, S. (2013). *Em2: A package for computing reading time measures for psycholinguistics* [R package version 0.9]. R package version 0.9. https://cran.r-project.org/src/contrib/Archive/em2/

Logačev, P., & Vasishth, S. (2016a). 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

Logačev, P., & Vasishth, S. (2016b). Understanding underspecification: A comparison of two computational implementations. *Quarterly Journal of Experimental Psychology*, *69*(5), 996–1012. https://doi.org/10.1080/17470218.2015.1134602

MacLeod, C. M., Gopie, N., Hourihan, K. R., K. L. Neary, & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671–685. https://doi.org/10.1037/a0018785

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

Mertzen, D., Paape, D., Dillon, B., Engbert, R., & Vasishth, S. (2023). Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data. *Glossa Psycholinguistics*. https://doi.org/10.5070/G60111266

Miller, G. A. (1962). Some psychological studies of grammar. *American Psychologist*, *17*, 748–762. https://doi.org/10.1037/h0044708

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 419–491). New York, NY, John Wiley; Sons.

Mitchell, D. C., Cuetos, F., & Zagar, D. (1990). Reading in different languages: Is there a universal mechanism for parsing sentences? In D. A. Balota, G. B. F. d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading*. Hillsdale, New Jersey, Erlbaum.

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue: ''Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1–5. https://doi.org/10.1016/j.jmp.2016.01.002

Ness, T., & Meltzer-Asscher, A. (2017). Working memory in the processing of long-distance dependencies: Interference and filler maintenance. *Journal of Psycholinguistic Research*, (46), 1353–1365. https://doi.org/10.1007/s10936-017-9499-6

Nicenboim, B., Schad, D. J., & Vasishth, S. (2023). *An introduction to bayesian data analysis for cognitive science* [Under contract]. CRC Press. https://vasishth.github.io/bayescogsci/book/

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*. https://doi.org/10.1111/cogs.12589

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? evidence from new data and a bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427. https://doi.org/10.1016/j.neuropsychologia.2020.107427

Nicol, J., & Antón-Méndez, I. (2009). Time and again: Theoretical perspectives on formal linguistics in honor of D. Terence Langendoen. In W. Lewis, S. Karimi, H. Harley, & S. Farrer (Eds.). Amsterdam, John Benjamins Publishing. https://doi.org/10.1075/la.135.10nic

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. https://doi.org/10.1038/nn.2886

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu Wolfsthurn, S. V. G., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory [Special Issue on Memory Models]. *Journal of Memory and Language*, *55*(4), 601–626. https://doi.org/10.1016/j.jml.2006.08.009

Paape, D., Vasishth, S., & Engbert, R. (2021). Does local coherence lead to targeted regressions and illusions of grammaticality? *Open Mind*, *5*, 42–58. https://doi.org/10.1162/opmi_a_00041

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

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory*, *21*(8), 904–915. https://doi.org/10.1080/09658211.2013.766754

Raaijmakers, J., Schrijnemakers, J., & Gremmen, F. (1999). How to deal with the "language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*(3), 416–426. https://doi.org/10.1006/jmla.1999.2650

Rabe, M. M., Vasishth, S., Hohenstein, S., Kliegl, R., & Schad, D. (2020). Hypr: An R package for hypothesis-driven contrast coding. *The Journal of Open Source Software*, *5*, 2134. https://doi.org/10.21105/joss.02134

Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2024). Seam: An integrated activation-coupled model of sentence processing and eye movements in reading. *Journal of Memory and Language*, *135*, 104496. https://doi.org/10.1016/j.jml.2023.104496

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rich, S., & Wagers, M. (2020). Semantic similarity and temporal contiguity in subject-verb dependency processing, In *33rd Annual CUNY Conference on Human Sentence Processing*.

Rouder, J. N., Haaf, J., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*, 102–113. https://doi.org/10.3758/s13423-017-1420-7

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. New York, Chapman; Hall, CRC Press.

Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates: Implications for expectation and memory-based accounts. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00403

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386. https://doi.org/10.1016/s1364-6613(02)01958-7

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

Schlueter, Z., Parker, D., & Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology*, *10*, 1002. https://doi.org/10.3389/fpsyg.2019.01002

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*(3), 196–201. https://doi.org/10.1198/000313002137

Smith, G., & Vasishth, S. (2020). A principled approach to feature selection in models of sentence processing. *Cognitive science*, *44 12*, e12918. https://doi.org/10.1111/cogs.12918

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation* (Vol. 13). John Wiley & Sons. https://doi.org/10.1002/0470092602

Stone, K., Nicenboim, B., Vasishth, S., & Roesler, F. (2023). Understanding the effects of constraint and predictability in ERP. *Neurobiology of Language*. https://doi.org/10.1162/nol_a_00094

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

Thornton, R., & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language*, *48*(4), 740–759. https://doi.org/10.1016/S0749-596X(03)00003-2

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

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403. https://doi.org/10.1016/j.cognition.2014.01.007

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. https://doi.org/10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166. https://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. https://doi.org/10.1016/j.jml.2011.05.002

Vasishth, S. (2023). Some right ways to analyze (psycho)linguistic data. *Annual Review of Linguistics*, *9*, 273–291. https://doi.org/10.1146/annurev-linguistics-031220-010345

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, *8*(10), 1–14. https://doi.org/10.1371/journal.pone.0077006

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

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794.

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. https://doi.org/10.1016/j.jml.2018.07.004

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

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

Ward, P., & Sturt, P. (2007). Linguistic focus and memory: An eye movement study. *Memory & Cognition*, *35*, 73–86. https://doi.org/10.3758/bf03195944