A cross-linguistic investigation of similarity-based interference in sentence comprehension

Daniela Mertzen

Department of Linguistics

Faculty of Human Sciences

University of Potsdam

Doctoral thesis submitted to the Faculty of Human Sciences at the University of Potsdam in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Linguistics

2021

Submitted: November 17, 2021
Defended: May 5, 2022

Supervisors:
Prof. Dr. Shravan Vasishth, University of Potsdam, Germany
Prof. Dr. Ralf Engbert, University of Potsdam, Germany

Reviewers:
Prof. Dr. Shravan Vasishth, University of Potsdam, Germany
Prof. Dr. Dan Parker, College of William & Mary, USA

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-55668
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-556685

Abstract

The aim of this dissertation was to conduct a larger-scale cross-linguistic investigation of similarity-based interference effects in sentence comprehension. Studies on interference can offer valuable insights into the mechanisms that are involved in long-distance dependency completion during sentence comprehension. These studies can address an important question, namely, which types of constraints are used in order to resolve long-distance dependencies during real-time sentence comprehension. The study of interference effects can also tell us something about the relative timing of the different types of constraints, and whether some types of constraints may take priority over others.

Many studies have investigated similarity-based interference effects in sentence comprehension, showing that syntactic and semantic information are employed during the formation of long-distance dependencies (e.g., Arnett & Wagers, 2017; Cunnings & Sturt, 2018; Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2011). Nevertheless, there are some important open questions in the sentence comprehension literature on interference that are critical to our understanding of the constraints involved in dependency resolution.

The work in this dissertation addresses the following open questions:

The first research question concerns the relative timing of syntactic and semantic interference in online sentence comprehension. Only few interference studies have investigated this question, and, to date, there is not enough data to draw conclusions regarding the time course of syntactic and semantic interference effects (Van Dyke, 2007; Van Dyke & McElree, 2011). One influential model of sentence parsing, the cue-based retrieval model, assumes by default that syntactic and semantic retrieval cues simultaneously guide long-distance dependency reso-

lution (Lewis & Vasishth, 2005). If both types of interference effects are observed during retrieval, that is, at the point at which a dependency must be completed, then this would support the view that both syntactic and semantic information are used simultaneously during dependency completion. If, however, syntactic effects precede semantic effects, then this might indicate that syntactic constraints have priority over semantic constraints (Cunnings & Sturt, 2014; Dillon et al., 2013; Sturt, 2003; Van Dyke & McElree, 2011). Our first cross-linguistic study explores the time course of syntactic and semantic interference in two eye-tracking reading experiments that implement the study design used in Van Dyke (2007). The first experiment tests English sentences with new stimuli to eliminate a potential confound in the materials of the original Van Dyke (2007) study. The second, larger-sample experiment investigates the relative timing of the two types of interference effects in German sentences. Overall, our data suggest that syntactic and semantic interference can arise simultaneously during retrieval.

The second research question concerns a special case of semantic interference in sentence comprehension: We investigate whether cue-based retrieval interference can be caused by semantically similar, sentence-external items which are not embedded in a syntactic structure. Our second interference study builds on a landmark study by Van Dyke and McElree (2006). The study design used in their study is unique in that it is able to pin down the source of interference as a consequence of cue overload during retrieval, when semantic retrieval cues do not uniquely match the retrieval target. Unlike most other interference studies, the design in Van Dyke and McElree (2006) is able to rule out an alternative explanation for the interference effect, that is, encoding interference. Encoding accounts postulate that it is not cue overload at the retrieval site but the erroneous encoding of similar linguistic items in memory that leads to interference (Lewandowsky et al., 2008; Oberauer & Kliegl, 2006). While Van Dyke and McElree (2006) reported cue-based retrieval interference from semantically-similar sentence-external distractors, the evidence for this effect was weak. A subsequent study did not show interference of this type (Van Dyke et al., 2014). Given these inconclusive findings, further research is necessary to investigate semantic cue-based retrieval interference.

To date, cue-based retrieval interference from sentence-external distractors has only been investigated in English. However, the theoretical question of interest is not constrained to English sentence processing. Our second study provides a larger-scale cross-linguistic investigation of cue-based retrieval interference from semantically similar, sentence-external items. Three larger-sample eye-tracking studies in English, German, and Russian tested cue-based interference in the online processing of filler-gap dependencies. Our study further extends the previous research by investigating interference in each language under varying task demands (Logačev & Vasishth, 2016; Swets et al., 2008). Overall, we see some very modest support for proactive cue-based retrieval interference in English. Unexpectedly, this was observed only under a low task demand. In order to consolidate this unexpected finding in English under a low task demand, more evidence needs to be accumulated. In German and Russian, there is some evidence against the interference effect. It is possible that interference is attenuated in languages with richer case marking.

In sum, the cross-linguistic experiments on the time course of syntactic and semantic interference from sentence-internal distractors support existing evidence of syntactic and semantic interference during sentence comprehension. Our data further show that both types of interference effects can arise simultaneously. Our cross-linguistic experiments investigating semantic cue-based retrieval interference from sentence-external distractors suggest that this type of interference may arise only in specific linguistic contexts.

Acknowledgements

First of all, I thank my advisor Shravan Vasishth for his consistent support, offering invaluable advice and guidance whenever needed. I also want to thank my second advisor Ralf Engbert for his support and his insightful feedback on this research project. I can't thank Brian Dillon enough for his mentorship and guidance, for hosting me at UMass Amherst twice, and for introducing me to corn mazes! It has been a true pleasure to work with such outstanding scientists over the last few years.

A very big thank you goes to the best lab group one could ask for, current and former members, as well as associated researchers: Dario Paape, Anna Laurinavichyute, Dorothea Pregla, Paula Lissón, Kate Stone, Garrett Smith, Himanshu Yadav, Pia Schoknecht, Titus von der Malsburg, Bruno Nicenboim, Daniel Schad, Serine Avetisyan, Sol Lago, João Verríssimo, Maximilian Rabe, and Lena Jäger (whose excellent classes during my time as a Master's student arguably got me into this mess in the first place).

For their help with data collection, I am grateful to our lab manager Johanna Thieke, interim lab manager Romy Leue, as well as Lisa Plagemann, Alexandra Lorson, Maria Korochkina, Elna Haffner, Marie de la Fuente, Luzie Ahrens, Hanna Eversmann, Chiara Tschirner (University of Potsdam), Ria Geguera, Christian Muxica, Austin Tero (UMass Amherst), and Maria Ignashina (HSE Moscow). A big thank you also to Annett Eßlinger and Michaela Schmitz for their support with all organisational issues.

Finally, I want to thank my family and friends, and most of all, my best friend and partner Phil who moved countries with me so I was able to pursue my studies, and cooked me many delicious meals to keep me healthy and happy throughout this fun, challenging and inspiring period of my life.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 317633480 — SFB1287, Project B03 (PIs: Shravan Vasishth and Ralf Engbert).

Table of Contents

List of Tables x

List of Figures xiii

1 Introduction 1

2 Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data 6

2.1 Introduction . . . 8

2.2 The present eye-tracking study . . . 13

2.2.1 Experimental design and materials . . . 14

2.2.2 Participants . . . 17

2.2.3 Procedure . . . 18

2.3 Predictions . . . 19

2.4 Statistical analyses . . . 21

2.5 Results . . . 23

2.5.1 Comprehension question accuracy . . . 23

2.5.2 Reading measure results . . . 24

2.6 Discussion . . . 32

2.7 General discussion . . . 34

2.7.1 The possible origins of effects at the pre-critical adverb . . . 37

2.8 Conclusion . . . 41

Appendix A Question responses 42

3 Is there cross-linguistic evidence for proactive cue-based retrieval interference in sentence comprehension? Eye-tracking data from English, German and Russian 45

3.1 Introduction . . . 47

3.1.1 Motivation for the present study . . . 52

3.1.1.1 Re-investigation of proactive interference . . . 52

3.1.1.2 A cross-linguistic investigation of proactive interference . . . 54
3.1.1.3 Proactive interference under different processing depths . . . 55
3.2 The current study . . . 57
3.2.1 Experimental design and materials . . . 57
3.2.1.1 The Memory load by Interference manipulation . . . 57
3.2.1.2 The Depth of Processing manipulation . . . 61
3.2.1.3 English vs. German and Russian materials . . . 63
3.2.1.4 Participants . . . 63
3.2.1.5 Procedure . . . 66
3.2.2 Predictions . . . 69
3.2.2.1 Predictions for the Depth of processing manipulation (complex vs. simple version) . . . 69
3.2.2.2 Predictions for the cross-linguistic comparison between German and Russian vs. English . . . 69
3.2.3 Statistical analyses . . . 69
3.2.3.1 Bayes factor analyses for model comparison . . . 71
3.2.4 Results . . . 73
3.2.4.1 Memory recall . . . 73
3.2.4.2 Comprehension question accuracy . . . 74
3.2.4.3 Reading times . . . 74
3.2.5 Discussion . . . 81
3.3 General Discussion . . . 83
3.4 Conclusions . . . 87

Appendix B Examples of German and Russian experimental materials 88
B.1 German stimuli . . . 88
B.2 Russian stimuli . . . 92

Appendix C Description of filler items 96
C.1 English . . . 96
C.2 German . . . 96
C.3 Russian . . . 97

Appendix D Perfect recall data 98
D.1 Comprehension question accuracy in perfect recall trials . . . 98

D.2 Raw by-condition reading times in perfect recall data 99
D.3 Reading time analyses for data with perfect recall 100

4 Synthesis of syntactic and semantic similarity-based interference effects **101**

5 Conclusion **111**

List of Tables

2.1 English example stimuli. Factor 1 (SUBJECTHOOD) manipulated whether the distractor (underlined) was a subject (+subj) or not a subject (−subj). Factor 2 (ANIMACY) manipulated whether the distractor was animate (+anim) or inanimate (−anim). 15
2.2 German example stimuli. Factor 1 (SUBJECTHOOD) manipulated whether the distractor (underlined) was a subject (+subj) or not a subject (−subj). Factor 2 (ANIMACY) manipulated whether the distractor was animate (+anim) or inanimate (−anim). 16
2.3 Sum contrast coding for effects of SUBJECTHOOD, ANIMACY, and their interaction. For the reading time measures, a SUBJECTHOOD or ANIMACY effect with a positive sign indicates slower reading times for +subject/+animate conditions, compared to −subject/−animate conditions. For the proportion of first-pass regressions and comprehension accuracy measures, a positive sign indicates a higher proportion of regressions, or a higher accuracy, for the +subject/+animate conditions, compared to the −subject/−animate conditions. 22
2.4 Comprehension question accuracies: Means of the posterior distributions with 95% credible intervals (CrIs) for the main effects of SUBJECTHOOD, ANIMACY and their interaction in English and German. 24

A.1 English: Given responses (in %) by condition for questions that targeted A) NP1 (*the attorney*), B) NP2 (*the secretary*), or C) (*the meeting/visitor*). 43
A.2 German: Given responses (in %) by condition for questions that targeted A) NP1 (*the journalist*), B) NP2 (*the colleague*), or C) (*the scandal/mafia boss*). 44

3.1 Example item (Van Dyke & McElree, 2006). 51

3.2 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. 58
3.3 English example items (complex and simple version). 59
3.4 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. 65
3.5 Shown are the location of the eye-tracking lab used for data collection for each study (English, German, Russian), and differences in the technical specifications of the in eye-tracking setup in the different labs. Chars = characters. 68
3.6 Contrast coding for effects of Load, Interference, and their interaction. 71
3.7 Guideline 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. 72

B.1 German example items (complex and simple version). 90
B.2 Russian example items (complex and simple version). 93

D.1 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. 100
D.2 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. 100

4.1 Similarity-based interference effects by eye-movement: Shown are the eye-tracking measures that were reported to show significant effects in experiments from 15 published interference studies. We show the proportion of significant effects for each reading measure, alongside the published experiments. The overall count of effects that were reported to be significant was 44. Of these, 11 of the 44 significant effects were reported in TFT (i.e., .25). The following reading measures are listed: (reg-cont) FFD = (regression-contingent) first fixation duration, FPRT = first-pass reading time, FPR = proportion of first-pass regressions out, RBRT = right-bounded reading time, Reg in = Regressions in, Cumu RT = cumulative reading times, RPD = regression-path duration, RRT = re-reading time, SPT = second-pass time, TFT = total fixation time. 103

List of Figures

2.1 Distribution of reading time differences between conditions from the Lewis & Vasishth (2005) model. Shown is the semantic interference effect nested within the −subject and within the +subject conditions. 20

2.2 By-condition means with 95% confidence intervals for question response accuracy in percent (%) in the English and the German experiment, respectively. Next to our results, we display the Van Dyke (2007) comprehension question accuracy results from their Experiment 2 (VD2007E2) and Experiment 3 (VD2007E3). The Van Dyke accuracy results were derived from the published paper. ±subj: distractor is (not) a subject, ±anim: distractor is (not) animate. 23

2.3 English: By-region plots: By-condition means with 95% confidence intervals (CIs) in A) first-pass reading times (FPRT), B) regression-path durations (RPD), and C) total fixation times (TFT). ±subj: distractor is (not) a subject, ±anim: distractor is (in)animate. . . . 25

2.4 German: By-region plots: By-condition means with 95% confidence intervals (CIs) in A) first-pass reading times (FPRT), B) regression-path durations (RPD), and C) total fixation times (TFT). ±subj: distractor is (not) a subject, ±anim: distractor is (in)animate. . . . 26

2.5 English (left panels): Posterior means with 95% credible intervals (CrIs) for the effects of SUBJECTHOOD, ANIMACY and their interaction at the pre-critical adverb (*frequently*), the critical verb (*complained*), and the post-critical region (*about the salary*). In German (right panels), the effect estimates are also shown at the pre-critical adverb (*tatsächlich*), the critical verb (*log,*), and the post-critical region (*um Informationen*). All values were back-transformed from the log-scale to the millisecond scale. FPRT = first-pass reading times, RPD = regression-path duration, TFT = total fixation times. Recall that a positive sign for the main effects of SUBJECTHOOD or ANIMACY indicate a reading time slowdown for +subject/+animate conditions, compared to −subject/−animate conditions. 28
2.6 First-pass regressions out (FPR) results in English (left panels) and German (right panels): Posterior means and 95% credible intervals (CrIs) for the effects of SUBJECTHOOD, ANIMACY and their interaction at the pre-critical adverb, the critical verb, and the post-critical region. All values were back-transformed from the log-odds scale to percentages. Recall that a positive sign for the main effects of SUBJECTHOOD or ANIMACY indicate a higher proportion of first-pass regressions out for +subject/+animate conditions, compared to −subject/−animate conditions. 29

3.1 Power estimates for an effect size $\mathcal{N}(30, 10)$ assuming various residual standard deviations (SD) (300, 250, 200, 150) for a different number of participants (40, 60, 80, 100, 120). Different shadings are merely for better distinguishability. The figure is adapted from Vasishth (2019). 53
3.2 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. 62
3.3 Example stimuli (schematic) for the English, German, and Russian experiment. 64

3.4 The dashed line shows that when the ν parameter of Stan's LKJ function is 0.5, less extreme correlation values are downweighted. The solid line shows that changing the ν parameter value to 2.0 downweights more extreme values. 71
3.5 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, GER = German, RU = Russian; (c) = complex version, (s) = simple version; VD&M = Van Dyke & McElree (2006); No int = No interference condition, Int = Interference condition. 74
3.6 By-condition mean comprehension question accuracies (in percent) with 95% confidence intervals. EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version. 75
3.7 By-condition mean reading times with 95% confidence intervals. (A) shows first-pass reading times (FPRT); (B) shows total reading times (TFT). EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version. 75
3.8 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. 77
3.9 Interaction estimates with 95% credible intervals from our study (first-pass reading times), compared with interaction estimate with 95% confidence interval from Van Dyke & McElree (2006) (self-paced reading times). EN = English, GER = German, RU = Russian, (c) = complex version, (s) = simple version. 78
3.10 Bayes factor results for Model 0 over Model 1 (BF01) in a) first-pass reading time 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 $\mathcal{N}(0, 0.1)$ 79

3.11 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 in English differs from German and Russian. The contrasts for the factor Language were specified using the R package `hypr` (Rabe et al., 2020), testing the null hypothesis H0: EN − (GER + RU)/2 = 0, i.e., whether the difference between the Load × Interference interaction in English (EN) vs. the interaction in German (GER) and Russian (RU) (averaged) is equal to zero. English was coded as +0.66, and German and Russian as −0.33. 80

D.1 By-condition mean comprehension question accuracies (in percent) with 95% confidence intervals in perfect recall trials. EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version 98
D.2 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, GER = German, RU = Russian; (c) = complex version, (s) = simple version. 99

4.1 The top panel shows posterior distributions of the syntactic interference effect for A) FPRT and SPR, B) RPD and SPR, C) TFT and SPR. The bottom panel shows the posterior distributions for the semantic interference effect. Dashed lines show the posterior means and 95% credible intervals. 106
4.2 Shown are the meta-analysis estimates (posterior means and 95% credible intervals) next to the individual effect estimates (means and 95% confidence intervals) of the pairwise comparisons that entered the meta-analysis. meta = meta analysis estimate, VD refers to the studies by Van Dyke et al., M refers to the experiments in this dissertation (Mertzen et al.), LoSyn/HiSyn = Low and High syntactic interference conditions, LoSem/HiSem = Low and High semantic interference conditions. EN = English, DE = German . . 107

4.3 Posterior distribution of the proactive semantic interference effect for FPRT and SPR. The interaction estimates from Van Dyke & McElree (2006) and Van Dyke et al. (2014), as well as the interaction estimates from our experiments in Chapter 3 (English complex, simple; German complex, simple; Russian complex, simple) entered the meta-analysis. Dashed lines show the posterior mean and 95% credible intervals. 109
4.4 Meta-analysis estimate (posterior mean and 95% credible interval) next to the interaction estimates (means and 95% confidence intervals for the Van Dyke et al. studies; posterior means and 95% credible intervals for the Mertzen et al. studies) that entered the meta-analysis. Meta = meta-analysis estimate, VanDyke06 = Van Dyke & McElree (2006), VanDyke14 = Van Dyke, Johns, & Kukona (2014), Mertzen20a refers to the studies in chapter 3, EN = English. DE = German, RU = Russian, c = complex version, s = simple version. 109

Chapter 1

Introduction

Reading sentences requires us to form dependencies between words that are non-adjacent. For instance, in sentence (1), the verb *jumped* and its subject *the squirrel* are separated by a relative clause. To comprehend this sentence, we must rapidly form the long-distance dependency between these two distal sentence elements.

(1) *The squirrel* that the fox spotted in the tree *jumped* from branch to branch.

Although sentences like (1) are perfectly well-formed, they can involve processing difficulty during real-time sentence parsing. This raises the question of what mechanisms are involved in the formation of long-distance dependencies. What information does the parsing system use to establish long-distance dependencies? And what are the imperfect properties of this mechanism that lead to processing difficulty during comprehension?

These questions have long been addressed in the psycholinguistic literature on sentence comprehension. According to the influential cue-based retrieval framework of parsing, linguistic items are encoded and maintained in memory until they need to be retrieved (Lewis & Vasishth, 2005; McElree, 2000; Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2011). This account specifies that the retrieval of an item is guided by a *cue-dependent* retrieval mechanism. It is assumed that this mechanism employs retrieval cues to re-activate, or retrieve, a target item from memory.

For example, in sentence (1), *the squirrel* needs to be retrieved at the verb *jumped*.

To retrieve the target item *the squirrel*, syntactic retrieval cues such as {*grammatical subject*}, and semantic retrieval cues such as {*animate*} are used.

Such a retrieval mechanism can lead to online processing difficulty, when there are other intervening items that are similar to the target of retrieval. It can also lead to a misretrieval of a distracting item, disrupting overall comprehension. In 1, *the fox* is such an item (a distractor). Like the features of the target, the features of the distractor also match the retrieval cues at the verb *jumped*. In the example above, both the target subject *the squirrel* and the distractor *the fox* are grammatical subjects as well as animate. When more than one item matches the retrieval cues, this can lead to a cue overload which, in turn, leads to processing difficulty during the retrieval process. This processing difficulty is referred to as *similarity-based interference*.

Many studies have reported similarity-based interference effects during online sentence comprehension (e.g., Arnett & Wagers, 2017; Cunnings & Sturt, 2014, 2018; Dillon et al., 2013; Fedorenko et al., 2006; Gordon et al., 2001, 2004; Gordon et al., 2006; Gordon et al., 2002; Jäger et al., 2020; Lowder & Gordon, 2014; Ness & Meltzer-Asscher, 2017; Nicenboim et al., 2018; Parker & Phillips, 2017; Schlueter et al., 2019; Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006, 2011; Wagers et al., 2009). Although similarity-based interference in sentence comprehension is a rather well-researched phenomenon, there are several open questions in the interference literature on sentence comprehension.

The goal of this dissertation is to address the following open questions:

The first research question concerns the time course of syntactic and semantic interference during online sentence comprehension. Can both types of information be used immediately and simultaneously, or do syntactic information precede semantic information? Additional data on this research question can shed more light on whether syntactic information may take precedence over semantic information during parsing—a question that has been discussed extensively in the work on ambiguity resolution in sentence parsing (e.g., Clifton et al., 2003; Ferreira & Clifton, 1986; Frazier, 1987; Frazier & Clifton, 1996; Frazier, 1979; Frazier & Rayner, 1982; Just & Carpenter, 1992; MacDonald et al., 1994; McRae et al., 1998; Tabor & Hutchins, 2004; Traxler & Frazier, 2008; Trueswell et al., 1993). The findings on this research question are conflicting (e.g., Frazier & Rayner, 1982; Trueswell et al., 1994).

In the interference literature on sentence comprehension of non-ambiguous structures, this question is also unresolved. Specifically, the results on the time course of semantic interference are inconclusive (e.g., Van Dyke, 2007; Van Dyke & McElree, 2011). In addition, the empirical data to date comes from studies on English sentence comprehension. The theoretical question of interest is, however, not constrained to English sentence processing. In chapter 2, we therefore present a cross-linguistic investigation of the time course of syntactic and semantic interference, using the Van Dyke (2007) design, in two larger-sample eye-tracking-while-reading experiments (English and German). Overall, our data show support for simultaneous syntactic and semantic interference effects.

The second open question concerns a special case of semantic interference: We investigate whether cue-based retrieval interference can be caused by semantically similar, sentence-external items. This work builds on a landmark study by Van Dyke and McElree (2006). The experimental design used in their study is unique in that it is able to ascribe any interference effects to a cue overload during retrieval, when semantic retrieval cues do not uniquely match the retrieval target. Unlike other studies on interference, the design in Van Dyke and McElree (2006) can rule out encoding interference as an alternative explanation for the interference effect. Encoding accounts posit that interference effects can result from faulty encoding of similar linguistic items in memory, as opposed to being the result of cue overload at the retrieval site (Lewandowsky et al., 2008; Oberauer & Kliegl, 2006).

Given the theoretical importance of this question, there are good reasons to re-investigate this claim. While Van Dyke and McElree (2006) reported cue-based retrieval interference from semantically-similar sentence-external distractors, the evidence for this effect was weak. First, for the reported effect, the $minF'$ statistic was not significant (Clark, 1973). Second, a subsequent study (Van Dyke et al.,

2014) that used the same experimental design as Van Dyke and McElree (2006) did not find this effect. The inconclusive findings require further investigation of cue-based retrieval interference from sentence-external distractors.

To date, this type of interference has also only been investigated in English (object-cleft constructions). If cue-based retrieval is a crucial part of human sentence processing, then it should be observed cross-linguistically. In chapter 3, we therefore present a larger-scale cross-linguistic test of cue-based retrieval interference from semantically similar, sentence-external items. Three larger-sample eye-tracking experiments test English, German, and Russian filler-gap dependencies. In addition to testing multiple languages, we examine interference under different processing depths (varying task demands encouraged deep versus superficial processing, Logačev and Vasishth, 2016; Swets et al., 2008).

Overall, our study finds only modest evidence for proactive cue-based retrieval interference in English. Surprisingly, the effect is observed only when superficial processing is encouraged by a simple task. However, German and Russian show some evidence against the interference effect. It is possible that the richer case marking in German and Russian, compared to English, can attenuate the interference effect. The unexpected finding in English would have to be consolidated in future studies.

One overarching aim of this dissertation was the re-investigation of the similarity-based interference effects discussed above. The re-investigation could be thought of as a conceptual replication of the previous work, that is, an investigation of the same research hypotheses, using different methods (Nosek & Errington, 2017). For example, for the experiments on interference from sentence-external distractors, if the parser considers semantically-similar, extra-sentential items as potential retrieval targets, then the effect should not only occur in English object-cleft sentences as tested in Van Dyke and McElree (2006). Instead, we would expect these results to be more generalizable (Nosek & Errington, 2017). For instance, we would also expect these effects in different syntactic structures (such as filler-gap dependencies), and when tested with a different reading method (such as eye-tracking).

We aim to test this here. In addition, we extend the interference research by investigating it cross-linguistically, and under different task demands.

Another overarching aim of the experiments in this dissertation was to obtain precise estimates of the interference effects under investigation. In chapter 3, we discuss that for similarity-based interference effects, prospective statistical power is relatively low with sample sizes commonly used in psycholinguistic studies. Underpowered studies result in either null results, or statistically significant results that are overestimates of the effect (Type M (magnitude) errors, Gelman and Carlin, 2014; in Vasishth et al., 2018, we discussed these issues in detail with an example from psycholinguistics). The estimate of the effect of interest in Van Dyke and McElree (2006) may have been such an overestimate. To obtain more precise effect estimates, we collected larger participant sample sizes than the previous studies. Chapters 2 and 3 report the effect estimates from each experiment separately. Finally, in chapter 4, we report meta-analysis estimates, separately for syntactic and semantic interference from sentence-internal distractors, and semantic interference from sentence-external distractors. These meta-analyses include the estimates from our experiments presented in chapters 2 and 3, as well as the results from the experiments by Van Dyke and colleagues. The results of the meta-analyses align with the results reported in 2 and 3.

Chapter 2

Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data

Daniela Mertzen, Dario Paape, Brian W. Dillon, Ralf Engbert, and Shravan Vasishth (2021a). Unpublished manuscript.

Abstract

A longstanding debate in the sentence processing literature concerns the time course of syntactic and semantic information in online sentence comprehension (Frazier & Rayner, 1982; Trueswell et al., 1993). Cue-based models of parsing assume by default that syntactic and semantic retrieval cues simultaneously guide long-distance dependency resolution (Lewis & Vasishth, 2005). When retrieval cues match multiple items in memory, this leads to similarity-based interference effects. Both syntactic and semantic interference effects have been shown empirically. However, the time course of these effects remains unclear (Van Dyke, 2007; Van Dyke & McElree, 2011). We report a cross-linguistic eye-tracking reading study (English, German) that uses the study design by Van Dyke (2007) to investigate the relative timing of syntactic and semantic cues in online sentence comprehension. The data suggest that syntactic and semantic interference can arise simultaneously. Interestingly, in both languages, the interference effects arise pre-critically. We discuss these findings in relation to cue-based theories of parsing.

2.1 Introduction

A longstanding debate in the literature on syntactic ambiguity resolution concerns the role of syntactic and semantic constraints during initial structure building. Consider, for instance, the garden-path sentences in (2), taken from Clifton et al. (2003):

(2) a. [[The man$_{\mathbf{NP}}$] [paid by the parents$_{\mathbf{RC}}$]] was unreasonable.

b. [[The ransom$_{\mathbf{NP}}$] [paid by the parents$_{\mathbf{RC}}$]] was unreasonable.

For these sentences, 'syntax-first' accounts of sentence processing would predict that comprehenders initially build an incorrect main clause analysis in which *paid* is analyzed as an active verb, regardless of the animacy status of the noun phrase (NP) *the ransom/the man*. This results in a garden-path effect at the by-phrase, where the structure must be reanalyzed as a reduced relative clause (Frazier, 1987; Frazier & Clifton, 1996; Frazier, 1979; Frazier & Rayner, 1982). Syntax-first models assume that in the earliest moments of structure building, only grammatical constraints play a role. Semantic information is used only in a subsequent processing stage to interpret the sentence.

By contrast, constraint-based accounts assume that syntactic and semantic constraints can be used in parallel (e.g., MacDonald et al., 1994; McRae et al., 1998; Tabor and Hutchins, 2004; Trueswell et al., 1993). These accounts predict that, in (2a), the inanimate NP *the ransom* should not be considered as a potential subject of *paid* due to its implausible interpretation, leading the parser towards the correct relative clause analysis, and thus eliminating the garden-path effect.

The evidence for these two competing accounts is mixed: In support of the syntax-first view, Ferreira and Clifton (1986) and Clifton et al. (2003) found that for sentences such as (2a,b), both conditions caused initial processing difficulty at the by-phrase, regardless of the animacy status of the NP. Animate conditions such as (2a) caused processing difficulty only in later sentence regions and in re-reading. These results are consistent with the hypothesis that syntactic constraints precede semantic constraints during real-time ambiguity resolution (see also Frazier and Rayner, 1982; Pickering and Traxler, 1998; Rayner et al., 1983; Traxler, 2002, 2005; Trueswell et al., 1993).

Meanwhile, a number of other studies have found support for the assumption that semantic information is used immediately, consistent with constraint-based models of parsing (Just & Carpenter, 1992; Tabor et al., 2004; Traxler & Frazier, 2008; Trueswell et al., 1994). For example, in sentences like (2), Trueswell et al. (1994) observed processing difficulty for animate, but not for inanimate conditions. Given the conflicting results, the debate on the time course of syntactic and semantic information remains unresolved.

The open question regarding the relative timing of syntactic and semantic influences on parsing is also of crucial interest outside of garden-path configurations. In the cue-based parsing framework, the time course of syntactic and semantic information has been studied in long-distance dependency resolution (e.g., Engelmann et al., 2019; Lewis & Vasishth, 2005; McElree, 2000; Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2011).

For example, in sentence (3), a dependency must be established between *the child* and *loved*:

(3) **The child** who the mother saw in the garden **loved** the rich chocolate cake.

Cue-based parsing assumes that the subject *the child* is encoded in memory, and subsequently retrieved at the matrix verb *loved*. This retrieval process is guided by retrieval cues, such as {*grammatical subject*} and {*animate*}, which are matched against the memory encodings of the nouns to seek out the correct target. In (3), the syntactic and semantic retrieval cues {*grammatical subject*} and {*animate*} do not only match the correct target *the child*, but also the intervening distractor NP *the mother*. Cue-based retrieval theory assumes that when the retrieval cues match multiple similar items in memory, it is more difficult for the processor to identify the correct target noun. The resulting processing difficulty is known as similarity-based interference. If a distractor partially matches the retrieval cues, this can also lead to a misretrieval of the distractor, which results in misinterpretation. For instance, in (3), *the mother* would be interpreted as the subject of *loved*.

The default assumption in the cue-based retrieval theory is that all retrieval cues are used immediately and simultaneously, similar to the parallel use of all information sources in constraint-based models. Thus, the default assumption would be that syntactic and semantic cues are used simultaneously during retrieval. However, the cue-based retrieval model could be implemented such that syntactic cues have precedence over other cues, for example, through cue-weighting (e.g., Dillon et al., 2013; Engelmann et al., 2019; Parker, 2019; Yadav, Paape, et al., 2021). While a stronger weighting for syntactic cues still assumes simultaneous use of retrieval cues, one implication is that effects of semantic interference manipulations should be weaker than those of syntactic interference manipulations.

The assumption of differential cue weighting can in principle be extended to also allow for differential "cue lag", such that syntactic cues are evaluated before semantic cues. This evaluation lag would be compatible with the proposal that syntactic cues may serve a "gating" function, ruling out syntactically mismatching chunks from being considered as retrieval targets in an early processing stage (Nicol & Swinney, 1989; Sturt, 2003; Van Dyke & McElree, 2011).

However, there are studies that suggest that non-syntactic cues may affect later processing, pointing towards a multi-stage processing architecture (Cunnings & Sturt, 2014; Lago et al., 2015; Sturt, 2003; Van Dyke, 2007; Wagers et al., 2009). To our knowledge, a multi-stage retrieval model has not been computationally implemented. In this paper, we present simulations from the cue-based retrieval model that show the predictions based on the default assumption that syntactic and semantic cues are used simultaneously.

Many empirical studies have reported interference from syntactically or semantically similar distractor items during retrieval (syntactic interference, e.g., Arnett and Wagers, 2017; Van Dyke, 2007; Van Dyke and Lewis, 2003, see also Glaser et al., 2013; semantic interference, e.g., Cunnings and Sturt, 2018; Glaser et al., 2013;

Gordon et al., 2004; Gordon et al., 2006; Laurinavichyute and von der Malsburg, 2021b; Lowder and Gordon, 2014; Rich and Wagers, 2020; Tabor et al., 2004; Van Dyke and McElree, 2011). However, only a small subset of studies has manipulated syntactic and semantic interference at the same time, which is necessary to uncover the relative timing of the two types of constraints.

One important study in this context is Van Dyke (2007), which manipulated the distractor's subjecthood (no subject, subject) and animacy (inanimate, animate) in a 2 × 2 repeated measures factorial design. In example (4) of their eye-tracking reading experiment (Experiment 2), the distractor *warehouse/neighbor* intervenes between the critical verb *was complaining* and the target subject *the resident*. In (4a, b), the distractor is not a grammatical subject, and should not cause syntactic interference. By comparison, in (4c, d), the distractor (like the target) is a grammatical subject and should therefore cause syntactic interference. Analogously, the animacy manipulation should not create semantic interference when the distractor noun is inanimate (e.g., *the warehouse* in 4a, c), but it should create semantic interference when it is animate (e.g., *the neighbor* in 4b, d).

A second eye-tracking experiment (Experiment 3), adapted the items of Experiment 2 by adding an adverbial phrase to the embedded clause to remove a potential confound in the stimuli of Experiment 2: reading time slowdowns may occur as a consequence of reading two adjacent verbs in syntactic interference conditions like (4c, d), and may not be associated with syntactic interference. Experiment 3 was thus an important check of the reading time patterns in Experiment 2.

(4) ...**the resident**

a. who was living [$_{PP}$ near the dangerous *warehouse*]
b. who was living [$_{PP}$ near the dangerous *neighbor*]
c. who said that [$_{NP}$ the *warehouse*] was dangerous
d. who said that [$_{NP}$ the *neighbor*] was dangerous

was complaining...

In both experiments, Van Dyke (2007) found reading time patterns consistent with syntactic and semantic interference effects. However, the effects were observed at different sentence regions. Reading time slowdowns consistent with syntactic interference were observed at the retrieval point, *was complaining*.[1] By contrast, reading time slowdowns consistent with semantic interference were observed only at the post-critical or sentence-final region.[2] This is at odds with the assumption that all cues are used immediately during retrieval. Surprisingly, Experiment 3 also found a reading time slowdown consistent with semantic interference at the pre-critical adverbial phrase in total fixation times. However, this effect was ascribed to a difference in between-condition plausibility observed in a pretest.

In a subsequent study, Van Dyke and McElree (2011) report a different time course for semantic interference in similar syntactic configurations. In this study, one eye-tracking experiment (experiment 1B) manipulated the animacy of intervening subject distractors. A second eye-tracking experiment (experiment 2B) manipulated the animacy of distractors that were in an object position, that is, the distractor did not additionally match the syntactic subject cue. In experiment 1B that had the subject distractors, semantic interference was observed during retrieval, at the critical verb. By contrast, experiment 2B which had the object distractors, showed no indication of a semantic effect at any sentence region. This pattern might suggest that semantic interference only occurs when the distractor additionally shares syntactic features with the target noun.

Overall, while there is some support for semantic interference effects in Van Dyke et al.'s studies, it is unclear whether semantic effects are immediate or not. Given the equivocal results of the above studies, more research is necessary on the timing of semantic and syntactic interference effects.

In the present work, we use the design of Van Dyke (2007) to investigate the time

[1]For the main effect in Experiment 3, the $minF'$ statistic was not significant (regression-path durations: p<.08; total fixation times: p<.09). $MinF'$ must be significant to claim that the effect can be generalized over subjects and items (Clark, 1973).

[2]Experiment 2 showed a pattern at the post-critical region that was consistent with an interaction in first-pass reading times, with longer reading times for animate compared to inanimate conditions, but only for conditions where the distractor was not a subject. In Experiment 3, a pattern consistent with semantic interference was reported at the sentence-final region. For the interaction and the main effect, the $minF'$ statistic was also not significant (p<.08, and p<.06, respectively).

course of syntactic and semantic constraints during retrieval. Our study uses new stimuli to eliminate a potential confound that was present in the materials of the Van Dyke (2007) study.

A further gap in the literature is that previous work on the topic has focused on English. It is vital to broaden the cross-linguistic scope of the investigation if the goal is to investigate theoretical claims that are not specific to English sentence processing.

In order to broaden the cross-linguistic scope of our research question concerning the time course of syntactic and semantic interference, we conducted a second eye-tracking study in German, using the Van Dyke (2007) design. The German study has another attractive property: it has a much larger sample size than many of the previous interference studies, which will provide more precise estimates of the interference effects.

If both syntactic and semantic interference effects are observed immediately at the retrieval point, this would support the view that all cues are used simultaneously during dependency completion. However, if only syntactic interference is observed during retrieval, this would speak in favor of syntactic information preceding semantic information during online dependency resolution, in line with the claim of syntax-first models. The timing of syntactic and semantic processing is usually discussed in dichotomous terms in the sentence processing literature, although it is possible that there is stochastic variability in the parsing system that leads to occasional simultaneous use of syntactic and semantic cues, and occasional staggering of cues.

2.2 The present eye-tracking study

Our study tested subject-verb dependencies similar to Van Dyke (2007). However, our design addresses a potential confound in the materials of the original study: In subject distractor conditions, such as sentence (4c, d), the critical verb region immediately followed two clause boundaries, while in the object distractor conditions, such as sentence (4a, b), it followed only one clause boundary. The processing of the additional clause boundary may create more difficulty, raising the possibility that the observed effects on the critical verb were complexity effects, and not the result of syntactic interference from subject distractors (this was previously discussed in Wagers, 2008, and Arnett and Wagers, 2017). We addressed this potential confound by adding an adverb, not in the embedded clause prior to the clause boundary like the original study, but as a pre-verbal modifier. Any clause boundary wrap-up effects should then be observed on the pre-critical adverb, and not on the critical verb.

2.2.1 Experimental design and materials

Our eye-tracking-while-reading experiments used a 2 × 2 repeated measures factorial design. As in the original Van Dyke (2007) design, syntactic interference was tested by manipulating Distractor subjecthood (−subject, +subject), and semantic interference was tested by manipulating Distractor animacy (−animate, +animate). Here, + denotes that the distractor is a subject/animate, and − means that it is not a subject/not animate. In the following we refer to the factors Subjecthood and Animacy to abbreviate Distractor subjecthood/animacy. The two factors were manipulated within-subjects and within-items, resulting in the four conditions shown in Table 2.1.

In all conditions, the critical dependency is between the verb *complained* (the point of retrieval) and its subject NP *the attorney* (the target of retrieval). Retrieval cues at the verb, e.g., {*grammatical subject*} and {*animate*}, always fully match the features +subject, +animate of the target NP.

The critical manipulation concerns the subjecthood and animacy of the distractor noun *meeting/visitor*. In the −subject conditions (a, b), the distractor is the direct object of the relative clause. Therefore, it does not match the retrieval cue {*grammatical subject*}. By contrast, in the +subject conditions (c, d), the distractor is the subject of the embedded subordinate clause, matching the {*grammatical subject*} cue. In the +animate conditions (b, d), the distractor matches the animacy

Table 2.1: English example stimuli. Factor 1 (SUBJECTHOOD) manipulated whether the distractor (underlined) was a subject (+subj) or not a subject (−subj). Factor 2 (ANIMACY) manipulated whether the distractor was animate (+anim) or inanimate (−anim).

It turned out that **the attorney**$^{+subj}_{+anim}$

a. −subject; −animate
whose secretary had forgotten about the important <u>meeting</u>$^{-subj}_{-anim}$ frequently

b. −subject; +animate
whose secretary had forgotten about the important <u>visitor</u>$^{-subj}_{+anim}$ frequently

c. +subject; −animate
whose secretary had forgotten that the <u>meeting</u>$^{+subj}_{-anim}$ was important frequently

d. +subject; +animate
whose secretary had forgotten that the <u>visitor</u>$^{+subj}_{+anim}$ was important frequently

complained$\{^{subj}_{anim}\}$ about the salary at the firm.

cue, while in the −animate conditions (a, c), it mismatches the animacy cue at the verb. Whenever there is a (partial) match between the retrieval cues at the verb and the features of the distractors, this should lead to similarity-based interference during the retrieval of the target *the attorney*.

Our stimuli contain an additional animate distractor (*secretary*) in the relative clause across all conditions. This second distractor intervenes between the critical verb and target (retroactive interference configuration). By contrast, in the Van Dyke (2007) study, an additional animate distractor *preceded* the target (proactive interference configuration).[3] Because previous research suggests that retroactive interference may be stronger than proactive interference (Van Dyke and McElree, 2011; see also Jäger et al., 2017), we included this additional intervening distractor to increase the strength of the interference manipulation (Nicenboim et al., 2018; Parker & Phillips, 2017).

The design of the German experiment closely matched that of the English experiment. Table 2.2 shows example stimuli. Here, a dependency must be established

[3]The terms proactive and retroactive interference from the memory literature were introduced to psycholinguistics in Lewis, 1996.

between the verb *log* (*'lied'*) and the target NP *der Journalist* (*'the journalist'*). As in the English example, the distractor is either the direct object of the embedded clause in conditions (a, b), or the grammatical subject as in conditions (c, d). As in the English study, the SUBJECTHOOD factor is crossed with ANIMACY such that the distractor is either inanimate (*Skandal*, *'scandal'*; a, c), or animate (*Mafiaboss*, *'mafia boss'*; b, d).

Table 2.2: German example stimuli. Factor 1 (SUBJECTHOOD) manipulated whether the distractor (underlined) was a subject (+subj) or not a subject (−subj). Factor 2 (ANIMACY) manipulated whether the distractor was animate (+anim) or inanimate (−anim).

Es stellte sich heraus, dass **der Journalist**$^{+subj}_{+anim}$,
It turned out that the journalist,

a. −subject; −animate

dessen Kollege	von	dem	grauenhaften	<u>Skandal</u>$^{-subj}_{-anim}$	berichtet	hatte,	tatsächlich
whose colleague	*of*	*the*	*gruesome*	*scandal*	*reported*	*had,*	*indeed*

b. −subject; +animate

dessen Kollege	von	dem	grauenhaften	<u>Mafiaboss</u>$^{-subj}_{+anim}$	berichtet	hatte,	tatsächlich
whose colleague	*of*	*the*	*gruesome*	*mafia boss*	*reported*	*had,*	*indeed*

c. +subject; −animate

dessen Kollege	berichtet	hatte,	dass der	<u>Skandal</u>$^{+subj}_{-anim}$	grauenhaft	war,	tatsächlich
whose colleague	*reported*	*had,*	*that the*	*scandal*	*gruesome*	*was,*	*indeed*

d. +subject; +animate

dessen Kollege	berichtet	hatte,	dass der	<u>Mafiaboss</u>$^{+subj}_{+anim}$	grauenhaft	war,	tatsächlich
whose colleague	*reported*	*had,*	*that the*	*mafia boss*	*gruesome*	*was,*	*indeed*

log$\{^{subj}_{anim}\}$, um Informationen zu erhalten.
lied, to obtain information.

Conditions a,b: 'It turned out that the journalist whose colleague had reported on the gruesome scandal/mafia boss in fact lied to obtain information.'
Conditions c,d: 'It turned out that the journalist whose colleague had reported that the scandal/mafia boss was gruesome in fact lied to obtain information.'

German is a particularly good test case to rule out the potential confound in the Van Dyke (2007) study: The obligatory punctuation at clause boundaries ensured that the pre-critical adverb unambiguously modifies the critical verb. In addition, the punctuation rules out the possibility that a semantic effect in −subject distractor conditions is a consequence of a syntactically illicit but locally coherent parse. For example, in sentence (4a), the direct object of the embedded clause (*the dangerous neighbor*) forms a locally coherent parse with the matrix verb (*was complaining*). A similar locally coherent parse could be built in our English example sentences (Table 2.1b: ... *about* [*the important visitor complained*]). However, these analyses are not syntactically licensed (Tabor et al., 2004).

We created 40 experimental items for each of the two experiments. The English study had 92 fillers and the German study had 90 filler items which can be found in the supplementary materials. All experimental sentences and half of the filler sentences were followed by a comprehension question. For experimental items, the questions targeted one of the three NPs in the sentences (e.g., *Who complained?* in Table 2.1). Each question had four response choices: one of the three NPs, or 'I don't know'. For instance, the example item in Table 2.1 had the response choices *an attorney, a secretary, a visitor* or *'?'* (*'I don't know'*). For −animate conditions, instead of the inanimate NP, the manipulated animate distractor from the +animate conditions was used as a response choice.

2.2.2 Participants

Our English study tested 61 participants.[4] These were mostly undergraduate students from the University of Massachusetts Amherst, MA, USA, who were reimbursed with 15 USD. The mean age was 19 years (range 18 to 28); 75% of participants identified as female and 25% as male.

For the German experiment, 121 participants were tested. Subjects were undergraduate students from the University of Potsdam, Germany, who were reimbursed either with 15 Euros or course credit for their participation. The mean age was 24 (range 18 to 50), and 76 % reported female gender and 24 % male gender. All participants had normal or corrected-to-normal vision and reported no history of language disorders.

[4]We aimed to test up to 120 participants but testing stopped due to COVID-19-related lab closures.

2.2.3 Procedure

All participants gave informed consent to take part in the present study. For the eye-tracking experiment, the participants were seated in front of a presentation monitor (1440×900 resolution) with their head placed in a headrest.

To record eye-movements, an EyeLink 1000 eye-tracker with a tower mount was used. After a calibration procedure, a short practice session familiarized participants with the task. For each trial, participants first read a sentence and then answered a comprehension question. Sentences were presented in one line on the computer screen in a monospaced font (Consolas) of size 16. The eye-to-screen distance was 64 cm such that 4.5 characters were within one degree of visual angle.

The items were presented according to a Latin square design such that each participant saw only one condition of each item, and the order of the items was randomized for each participant. The response choices for comprehension questions were displayed in the center top, left, right and bottom of the screen. The 'I don't know' choice was always presented in the same location at the bottom of the screen. The presentation location of the other three response choices was randomized.

A break was offered halfway through the experiment, and participants were invited to take additional breaks whenever needed. After each break, a re-calibration was performed.

As the German experiment was tested in a different lab, the setup was not identical to the setup of the English study: an EyeLink 1000 Plus[5] was used to conduct monocular tracking of the right eye. The monitor resolution was 1920×1080. The German were presented in font size 14, and the eye-to-screen distance was 56 cm resulting in 2.4 characters within one degree of visual angle.

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

2.3 Predictions

Cue-based retrieval theory predicts a reading time slowdown for conditions with a +subject distractor compared to conditions with a −subject distractor. Such a main effect of SUBJECTHOOD would indicate syntactic interference. A reading time slowdown is also predicted for +animate distractor conditions compared to −animate distractor conditions. This main effect of ANIMACY would suggest semantic interference. A default assumption of the cue-based retrieval model is that retrieval cues are weighted equally (Lewis & Vasishth, 2005). Under this assumption, we would expect the reading time slowdowns for syntactic and semantic interference to occur simultaneously at the point of retrieval (the critical verb). If both syntactic and semantic interference occur at the critical verb in the same reading measure(s), this would be consistent with a simultaneous use of syntactic and semantic retrieval cues. By contrast, if only syntactic interference is observed at the critical verb, but semantic interference is only observed at the post-critical region, or not at all, this would favor syntax-first accounts of sentence processing.

We computed quantitative predictions from the Lewis and Vasishth (2005) model of cue-based retrieval for the current design, following the procedure and parameter settings used by Vasishth (2020). For the simulation, the model was set up to use three equally weighted cues for the retrieval at the verb: ±*subject*, ±*animate*, and ±*same_clause*. The addition of the ±*same_clause* cue is necessary to identify the correct subject. The cue serves as a stand-in for the additional information—besides the structural position of the NP and its animacy—the parser uses to achieve correct retrieval. With only two cues, in the +*subject*, +*animate* conditions, the model would otherwise predict a race between the target NP and the distractor NP (e.g., Jäger et al., 2020), with each being retrieved 50% of the time, given that both would be perfect matches.

The top panel of Figure 2.1 shows the predicted differences between the -*subject*, +*animate* and the −*subject*, −*animate* condition (one-match versus no-match), while the bottom panel shows the predicted differences between the +*subject*,

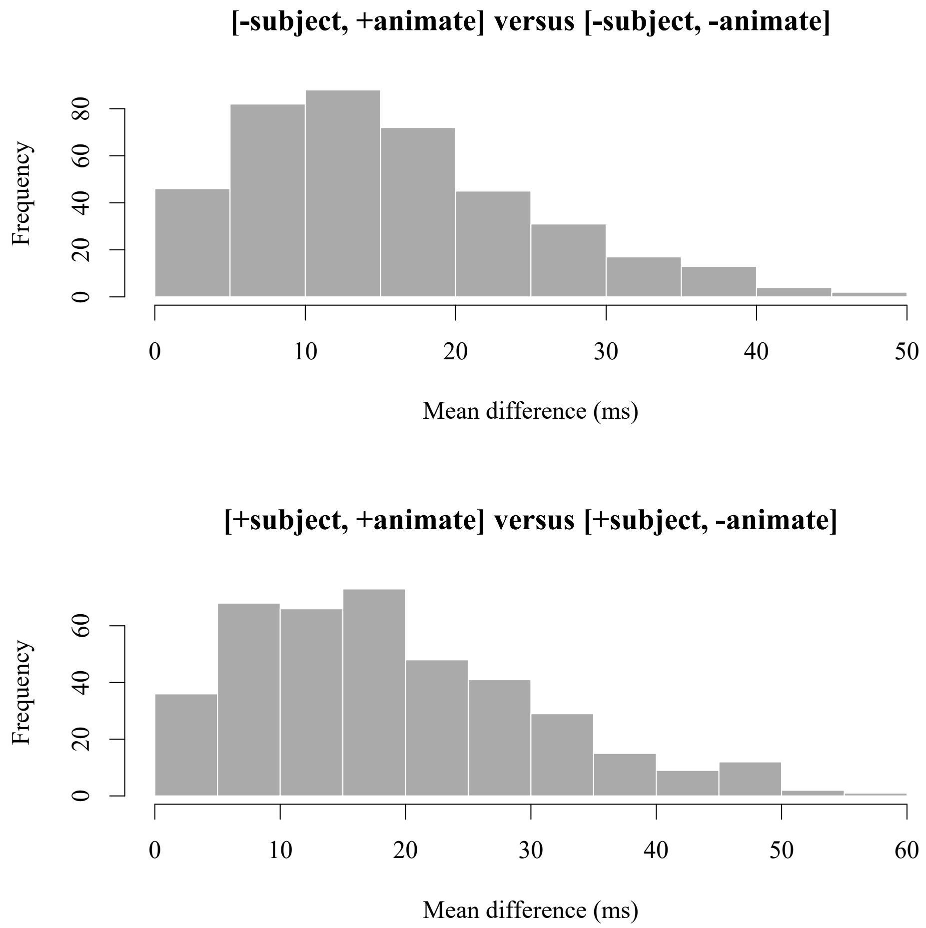


Figure 2.1: Distribution of reading time differences between conditions from the Lewis & Vasishth (2005) model. Shown is the semantic interference effect nested within the −subject and within the +subject conditions.

+*animate* and the +*subject*, −*animate* condition (two-match versus one-match). Given the assumed equal weighting of the cues, the −*subject*, +*animate* condition should show the same pattern as the +*subject*, −*animate* condition. The respective 95% ranges of the predicted mean differences are [2, 38] ms for no-match versus one-match and [3, 47] ms for two-match versus one-match. As the simulation results show, the model predicts a small superadditive interaction of the two interference types on the raw millisecond scale, due to the heavy right tail of the generated distribution. However, the superadditivity disappears when the differences are computed on the logarithmic scale, which we do for the empirical data. With activation noise set to 0.2, the model predicts zero misretrievals in the no- and one-match conditions, and 3% misretrievals in the two-match condition.

2.4 Statistical analyses

Following Van Dyke (2007), we analyzed three reading time measures (first-pass reading times, regression-path durations, and total fixation times), in addition to proportion of first-pass regressions. First-pass reading times (FPRT) include the sum of all fixations on a region n before a forward or backward saccade is launched. Regression-path durations (RPD) consist of the sum of all first-pass fixation durations on region n, including any fixation durations that result from regressions out of region n, until n is left to the right. Total fixation time (TFT) is defined as the sum of all fixations that occurred during the first pass and during re-reading of a region n. The proportion of first-pass regressions (FPR) measure is defined as the proportion of regressive saccades out of region n during the first-pass (Logačev & Vasishth, 2013; Rayner, 1998). FPRT, RPD, and TFT encapsulate so-called early and late reading processes, while the latter (FPR) allows us to quantify the cost of the manipulation on regressive eye movements which have been said to reflect integration difficulty of a word (Clifton et al., 2007).

Bayesian linear mixed effects models were fit to the reading time data (e.g., Gelman et al., 2014; Kruschke, 2015), using the probabilistic programming language Stan (Carpenter et al., 2017) in the programming environment R (R Core Team, 2019). A log-normal likelihood was assumed for the reading time data. The models included the factors Subjecthood (−subject, +subject), Animacy (−animate, +animate), and the interaction as fixed effects. These were sum-coded (Table 3.6). Subjects and items were specified in the models as random effects, with full variance-covariance matrices.

We specified mildly informative prior distributions for the model parameters. These priors are regularizing, i.e., they downweight values that are a priori very unlikely. A regularizing LKJ prior was also specified for the correlation matrix of

Condition	Subject	Animacy	Interaction
−subject, −animate	−0.5	−0.5	+0.5
−subject, +animate	−0.5	+0.5	−0.5
+subject, −animate	+0.5	−0.5	−0.5
+subject, +animate	+0.5	+0.5	+0.5

Table 2.3: Sum contrast coding for effects of SUBJECTHOOD, ANIMACY, and their interaction. For the reading time measures, a SUBJECTHOOD or ANIMACY effect with a positive sign indicates slower reading times for +subject/+animate conditions, compared to −subject/−animate conditions. For the proportion of first-pass regressions and comprehension accuracy measures, a positive sign indicates a higher proportion of regressions, or a higher accuracy, for the +subject/+animate conditions, compared to the −subject/−animate conditions.

the random effects variance-covariance matrix (Lewandowski et al., 2009). The shape parameter ν (nu) of the LKJ prior was set to 2. This means that extreme values like ±1 are downweighted.

Each model was run with four chains and 4000 iterations each. The first 2000 of these served as warm-up iterations, that is, they were not used for inference. $\hat{R}$-diagnostic and trace plots were inspected to check model convergence (Gelman et al., 2014).

We report the 95% credible intervals (CrIs) of the posterior distributions, which can be interpreted as the interval that contains the true parameter with 95% probability, given the data and the model. All estimates are reported on the millisecond scale; these were back-transformed from the log-scale (Nicenboim et al., 2021; Vasishth et al., 2018).

For proportions of first-pass regressions and question response accuracies, maximal hierarchical logistic models were fit with regularizing priors. Four chains with 4000 iterations were specified for each model. For first-pass regression- and response accuracy results, we also report the 95% credible intervals.

2.5 Results

2.5.1 Comprehension question accuracy

Figure 2.2 shows the raw by-condition accuracies for both our English and German experiments. Next to the accuracies from our study, we display the comprehension question accuracies of Van Dyke (2007)'s study (Experiments 2 and 3), which used similar sentences. Compared to Van Dyke's experiments, our experiment shows overall lower comprehension accuracies. This is likely because of the different question types between the two studies. In the Van Dyke (2007) experiments, the questions had a cloze format with two response choices (the target and one of the animate distractor NPs), whereas our questions had three NPs as response choices, in addition to the 'I don't know' option. In Appendix A, we display the responses by condition in both languages.

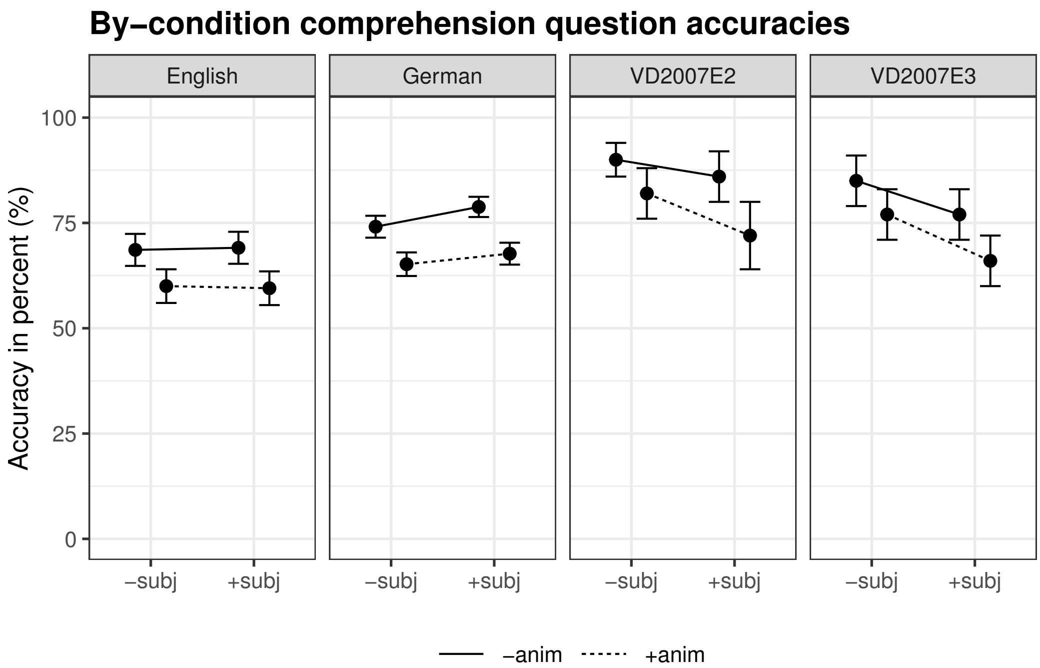


Figure 2.2: By-condition means with 95% confidence intervals for question response accuracy in percent (%) in the English and the German experiment, respectively. Next to our results, we display the Van Dyke (2007) comprehension question accuracy results from their Experiment 2 (VD2007E2) and Experiment 3 (VD2007E3). The Van Dyke accuracy results were derived from the published paper. ±subj: distractor is (not) a subject, ±anim: distractor is (not) animate.

Table 2.4 shows the English and German results for the statistical analysis of the accuracy data. In English, the most plausible values for the effect of SUBJECTHOOD are centered around zero. In German, the 95% CrI for the SUBJECTHOOD effect ranges from 1% to 8%, suggesting a somewhat higher comprehension accuracy when the distractor is also a subject. This contrasts previous results on English in Van Dyke (2007) that found lower accuracies for subject distractor conditions.

In English, the main effect of ANIMACY has 95% CrI [−15, −6]%, suggesting that comprehension question accuracy is lower when the manipulated distractor is animate compared to when it is inanimate. Similarly, in German, the estimate ranges from −16% to −7%, consistent with a lower accuracy for animate distractor conditions, i.e., semantic interference. This pattern for semantic interference was also reported in Van Dyke (2007).

	Comprehension question accuracy	
	English	**German**
Fixed effects	posterior mean [95% CrI]	posterior mean [95% CrI]
SUBJECTHOOD	0% [−6, 6]%	4% [1, 8]%
ANIMACY	−11% [−15, −6]%	−11% [−16, −7]%
INTERACTION	−1% [−6, 4]%	−2% [−5, 1]%

Table 2.4: Comprehension question accuracies: Means of the posterior distributions with 95% credible intervals (CrIs) for the main effects of SUBJECTHOOD, ANIMACY and their interaction in English and German.

2.5.2 Reading measure results

Figures 2.3 (English) and 2.4 (German) display the raw by region and condition mean reading times with their 95% confidence intervals in first-pass reading times, regression-path durations, and total fixation times.

Recall that the syntactic manipulation necessitates two embedded clauses in the +subject conditions, compared with one embedded clause in the −subject conditions. In previous work, this difference potentially confounded the critical region because a slowdown on the critical verb could have been a complexity effect, rather

English:

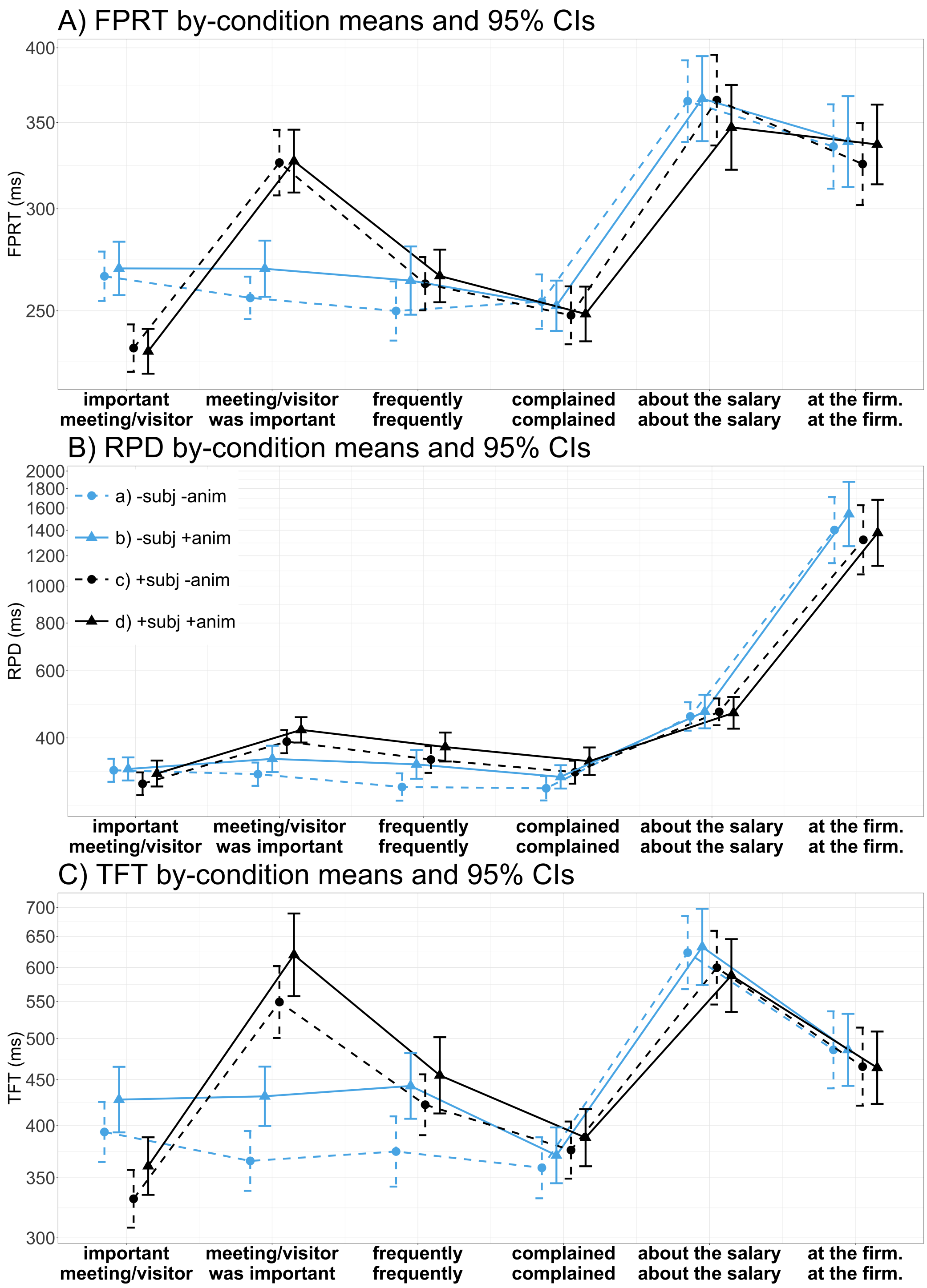


Figure 2.3: English: By-region plots: By-condition means with 95% confidence intervals (CIs) in A) first-pass reading times (FPRT), B) regression-path durations (RPD), and C) total fixation times (TFT). ±subj: distractor is (not) a subject, ±anim: distractor is (in)animate.

German:

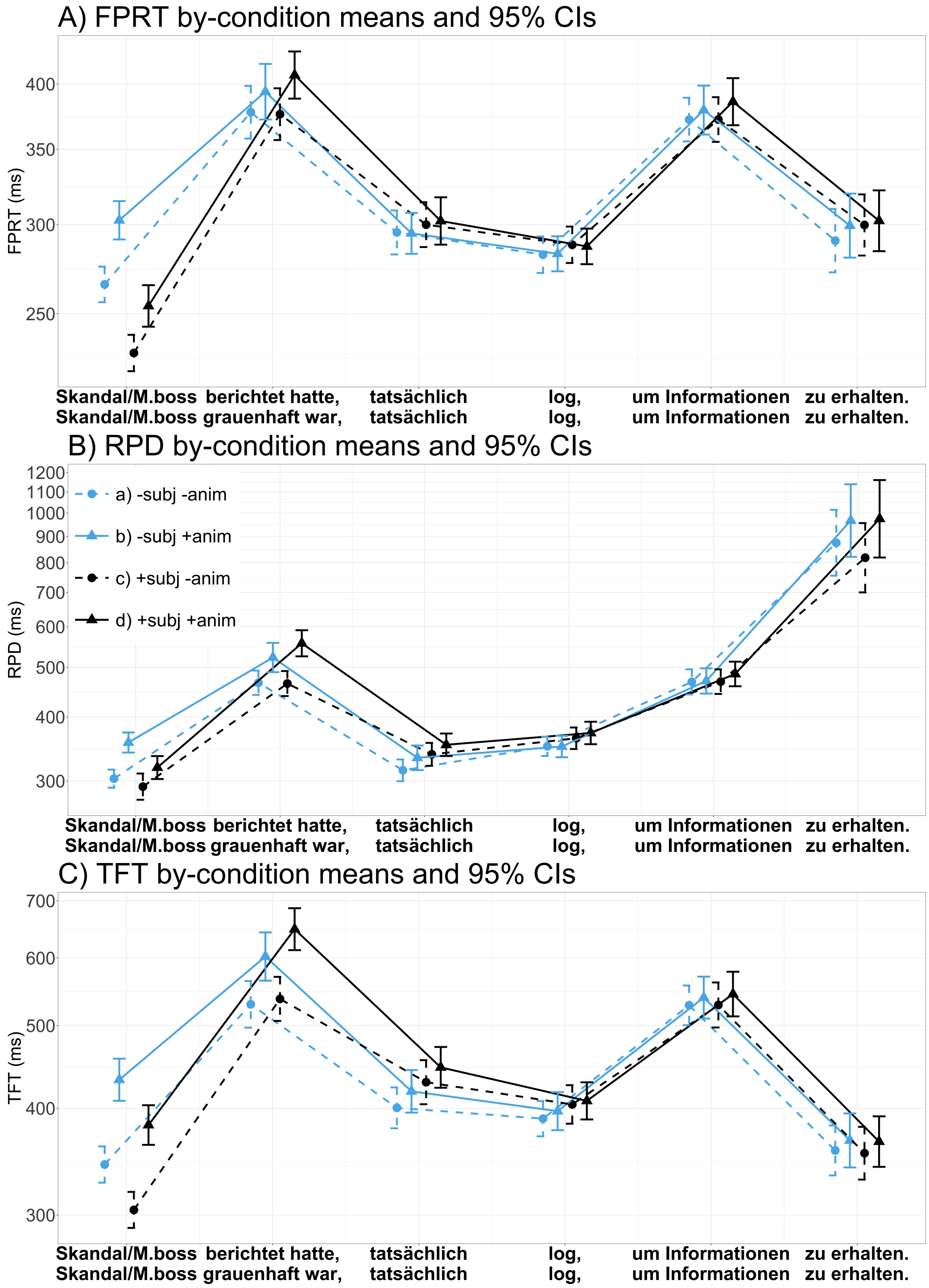


Figure 2.4: German: By-region plots: By-condition means with 95% confidence intervals (CIs) in A) first-pass reading times (FPRT), B) regression-path durations (RPD), and C) total fixation times (TFT). ±subj: distractor is (not) a subject, ±anim: distractor is (in)animate.

than similarity-based interference (Van Dyke, 2007). In our study, a pre-critical adverb was added such that any complexity effect is more likely to be observed on this adverb, rather than the critical verb. To check this, we conducted an additional analysis of the pre-critical region.

Figure 2.5 shows the English and German effect estimates at the pre-critical, critical, and post-critical regions, for first-pass reading times (FPRT), regression-path durations (RPD), and total fixation times (TFT). Figure 2.6 shows the results for first-pass regressions out (FPR).

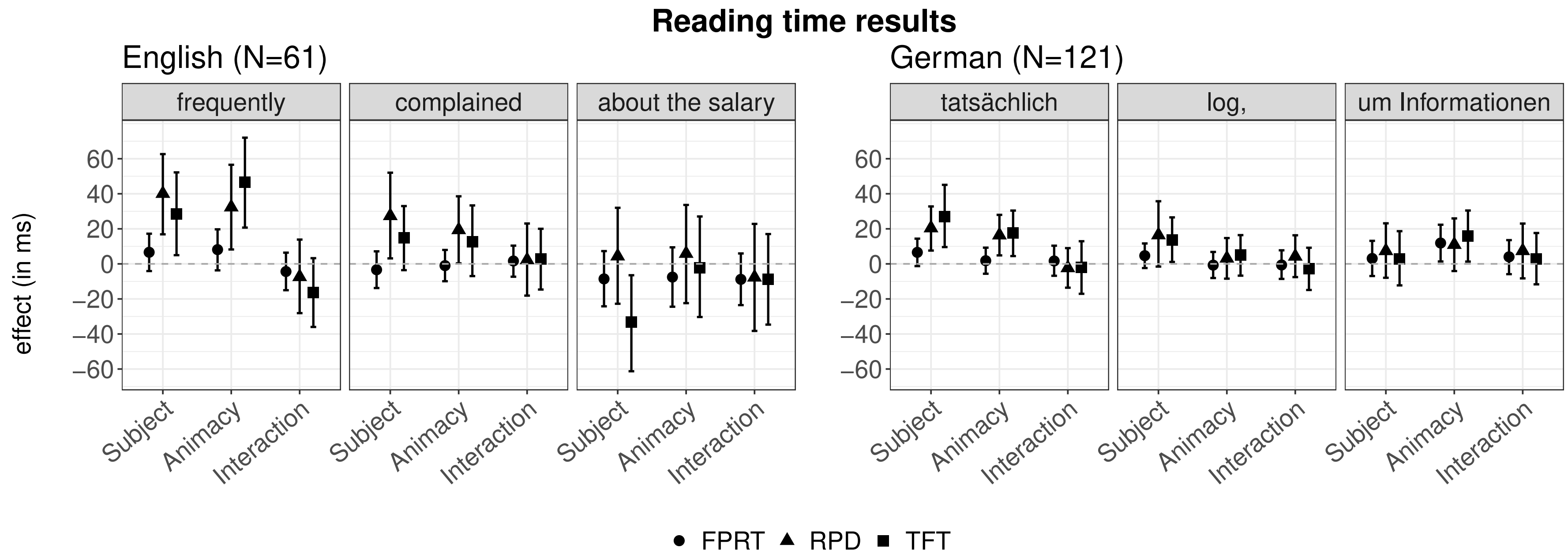


Figure 2.5: English (left panels): Posterior means with 95% credible intervals (CrIs) for the effects of SUBJECTHOOD, ANIMACY and their interaction at the pre-critical adverb (*frequently*), the critical verb (*complained*), and the post-critical region (*about the salary*). In German (right panels), the effect estimates are also shown at the pre-critical adverb (*tatsächlich*), the critical verb (*log,*), and the post-critical region (*um Informationen*). All values were back-transformed from the log-scale to the millisecond scale. FPRT = first-pass reading times, RPD = regression-path duration, TFT = total fixation times. Recall that a positive sign for the main effects of SUBJECTHOOD or ANIMACY indicate a reading time slowdown for +subject/+animate conditions, compared to −subject/−animate conditions.

Figure 2.6: First-pass regressions out (FPR) results in English (left panels) and German (right panels): Posterior means and 95% credible intervals (CrIs) for the effects of Subjecthood, Animacy and their interaction at the pre-critical adverb, the critical verb, and the post-critical region. All values were back-transformed from the log-odds scale to percentages. Recall that a positive sign for the main effects of Subjecthood or Animacy indicate a higher proportion of first-pass regressions out for +subject/+animate conditions, compared to −subject/−animate conditions.

Pre-critical adverb

English For the English experiment, the most plausible values of the main effect of SUBJECTHOOD largely have a positive sign (FPRT 95% CrI [−4, 17] ms, RPD [17, 63] ms, TFT [5, 52] ms). In FPR, the main effect of SUBJECTHOOD has 95% CrI [3, 11]%. The effects with a positive sign indicate reading time slowdowns and higher proportion of regressions out from the pre-critical region in the +subject distractor conditions.

The main effect of ANIMACY indicates a reading time slowdown and more regressions in +animate distractor conditions (95% CrI FPRT [−4, 20] ms, RPD [8, 57] ms, TFT [21, 72] ms, FPR [2, 8]%). Nested comparisons in RPD show that in +subject conditions, the 95% CrI is [4, 66] ms, and in −subject conditions, it is [−7, 53] ms. This slowdown would be consistent with a semantic interference effect. However, it is unexpected because it occurs before the critical verb.

German In German, we see a similar pattern to English. The main effect of SUBJECTHOOD has a positive sign, indicating a reading time slowdown, and more regressions, in +subject conditions (95% CrI FPRT [−1, 14] ms, RPD [8, 33] ms, TFT [10, 45] ms, FPR [0, 4]%).

The effect of ANIMACY also has a positive sign, suggesting slower reading times and a higher proportion of first-pass regressions in +animate conditions (95% CrI RPD [5, 28] ms, TFT [4, 30] ms, FPR [1, 4]%). Nested comparisons for RPD show that in +subject conditions, the 95% CrI is [3, 34] ms, and in −subject conditions, it is [−3, 29] ms. The FPRT estimate is centered around zero. As in English, the reading time patterns in RPD, TFT and FPR are compatible with semantic interference effects, although the critical retrieval was not expected at this region.

Critical verb region At this region, the target NP should be retrieved. Hence, syntactic and semantic interference effects were predicted to occur at this region if both types of cues are used simultaneously.

English The effect of SUBJECTHOOD has the following 95% CrIs: RPD [3, 52] ms, TFT [−4, 33] ms, and FPR [1, 8]%. This suggests a slowdown in reading times and a higher proportion of regressions in the +subject distractor conditions, consistent with syntactic interference.

For the main effect of ANIMACY, the RPD estimate has 95% CrI [0, 39] ms. Nested comparisons in RPD showed that in +subject conditions, the 95% CrI is [−9, 40] ms, and in −subject conditions, the 95% CrI is [−10, 48] ms. The effect of ANIMACY in TFT ranges from −7 ms to 33 ms. In FPR, the interval ranges from −1% to 5%, suggesting more regressions out from the critical region when the distractor is animate. These results would be consistent with a semantic interference effect.

German In German, similar to English, the most plausible values of the effect of SUBJECTHOOD largely have a positive sign (95% CrI FPRT [−2, 12] ms, RPD [−2, 36] ms, TFT [1, 27] ms). This slowdown for +subject distractor conditions is consistent with a syntactic interference effect. There is also a somewhat higher proportion of regressions for +subject compared to −subject distractor conditions (95% CrI FPR [−1, 5]%). However, unlike English, the most plausible values of the effect of ANIMACY are centered around zero in all reading measures.

Post-critical region

English In the English experiment the effect of SUBJECTHOOD in TFT has a negative sign (95% CrI [−61, −6] ms). This is consistent with a reading time speedup for +subject conditions. For the other measures, the most plausible values of the effect of SUBJECTHOOD are centered around zero. Similarly, the effect estimates for the main effect of ANIMACY are centered on zero. That is, there is no indication of a reading time slowdown for +animate distractor conditions at this region.

German In the German experiment, the reading time patterns differ from those in the English experiment. In all reading measures, the estimated effect of SUB-

JECTHOOD is centered on zero. However, the main effect of ANIMACY has a positive sign (95% CrI FPRT [1, 22] ms, RPD [−4, 26] ms, TFT [1, 30] ms). This suggests that there is a reading time slowdown for + animate conditions at the post-critical region in German. Nested comparisons in RPD show that in +subject conditions, the 95% CrI is [−18, 25] ms, and in −subject conditions, it is [−4, 41] ms.

2.6 Discussion

We begin by discussing the effects at the critical and post-critical region, as effects were expected to occur in those regions. In both languages, either the critical verb or the spillover region showed effects consistent with cue-based theories' predictions of syntactic and semantic interference effects during retrieval. In the English experiment, the critical verb region shows a reading time pattern that is consistent with both types of interference, that is, a slowdown in reading times and more first-pass regressions when the distractors are grammatical subjects, and when they are animate. Unexpectedly, in the post-critical region, the +subject distractor conditions showed a reading time speedup. Speculatively, this may reflect a recovery from the processing difficulty at the previous regions (Paape et al., 2018).

In German, the critical verb showed a reading time slowdown in +subject conditions that is consistent with a syntactic interference effect. In contrast to English, this region did not show any indication of a semantic interference effect. However, at the post-critical region, a reading time slowdown was observed in +animate distractor conditions that is consistent with semantic interference.

The earlier syntactic and the later semantic effect in German could, in principle, be suggestive of syntactic effects preceding semantic effects, which would be consistent with syntax-first theories. However, it is possible that this categorical distinction is too strict. There is no principled reason why the time course of processing of syntactic and semantic information should differ in English compared to German. Therefore, it is possible that the later occurrence of the semantic effect in German is due to processing spillover: The critical retrieval may be triggered at the verb with both cues, but semantic integration may be completed only at the post-verbal region.

What assumptions should cue-based retrieval models make given these data that show an immediate semantic effect in English, but a delayed semantic effect in German? It is, in principle, possible in these models to implement a small lag in semantic compared to syntactic effects. Syntax may take precedence over semantics in settings where syntactic structure needs to be built prior to evaluating the semantic fit of the resulting parse. For example, consider the sentence (5) from Van Gompel et al. (2000).

(5) While the guests were eating plates were brought in.

Here, a transitive interpretation of the verb *eating* is temporarily syntactically possible, but implausible. In order to evaluate the implausible semantics of this parse (*eating plates*), the parse first needs to be constructed. Van Gompel et al. (2000) postulate that while non-syntactic information can be used rapidly, the logical constraint of semantic plausibility information lagging behind in some settings can result in processing difficulty. While a similar logic could be applied to our sentences—first retrieve a subject, then evaluate its semantic fit—this would still not explain the difference between English and German.

An alternative explanation is that the coupling between the parser and oculomotor control is not perfect, such that the post-critical region is sometimes fixated before the critical word has been fully processed (Vasishth et al., 2013). This can lead to effects at the spillover region, which are frequently observed in reading studies on interference (e.g., Lago et al., 2015; Pearlmutter et al., 1999; Sturt, 2003; Van Dyke, 2007). In our study, the semantic effect on the post-critical region already occurs in first-pass reading times. Because the post-critical slowdown arises in the earliest moments of processing, this could reflect a spillover effect from the critical verb which has not been fully processed. Nevertheless, it is puzzling that the effect would spill over to the subsequent region only in German but not in English.

Although the spillover explanation is speculative, there is another indicator that semantic effects are not generated later than syntactic effects in German, namely, that a semantic interference effect was also observed on the pre-critical adverb. Recall that in both languages, the syntactic and the semantic manipulation caused reading time slowdowns and a higher proportion of first-pass regressions on the pre-verbal modifier. This could be indicative of interference arising prior to the critical region. If so, syntactic interference does not precede semantic interference. Before we turn to the effects at the pre-critical region, we will summarize the findings at the critical regions where interference effects were predicted to occur.

2.7 General discussion

Our study asked whether syntactic and semantic interference effects arise simultaneously during online dependency formation, as predicted by cue-based retrieval theories. To evaluate the cross-linguistic support for syntactic and semantic interference during retrieval, two eye-tracking experiments tested similarity-based interference in English and German. Both languages were tested with the same experimental method and design, as well as similar syntactic constructions, namely, subject-verb dependencies that do not involve number agreement (henceforth, non-agreement subject-verb dependencies, Jäger et al., 2017).

In both languages, reading times showed patterns that are consistent with syntactic and semantic interference effects. Our data thus add to the large body of evidence on syntactic and semantic interference effects during online dependency resolution (e.g., Arnett and Wagers, 2017; Cunnings and Sturt, 2018; Dillon et al., 2013; Laurinavichyute and von der Malsburg, 2021b; Lowder and Gordon, 2014; Tabor et al., 2004; Van Dyke, 2007; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2011).

Our English data showed reading time slowdowns and a higher proportion of first-pass regressions due to syntactic and semantic interference directly at the critical verb. Similarly, the German data also showed reading time slowdowns and more regressions immediately at the critical verb, but only for syntactic interference. Semantic interference was observed at the post-verbal region, in the form of a reading time slowdown.

Our study removed a confound present in the materials of Van Dyke (2007) and Van Dyke and Lewis (2003), namely, that the observed syntactic interference finding could also be explained by a sentence complexity effect due to two versus one clause boundaries across conditions. Our experiments, which removed this confound by adding a pre-verbal modifier region, support the Van Dyke (2007) claim of syntactic interference.

This being said, our findings regarding the time course of semantic interference differ from those of Van Dyke (2007). In Van Dyke's study, semantic effects were seen on the sentence-final region and were attributed to sentence wrap-up. The offline comprehension data also suggested an effect of semantic interference. While our English study suggests that semantic interference can arise during retrieval, it is possible that semantically similar distractors continue to affect processing and have a lasting detrimental effect on overall comprehension. Our offline accuracy data points to more frequent misinterpretations when sentences have several semantically similar nouns. This suggests that despite the prolonged reading times, and higher proportion of regressions, comprehenders do not always arrive at the correct sentence interpretation.

Our findings on semantic interference partially corroborate the effects observed in a further study by Van Dyke and colleagues (Van Dyke & McElree, 2011). Both in our English study and in the study of Van Dyke and McElree (2011), semantic interference was observed at the retrieval point when distractors were animate, that is, when they matched the semantic cue. However, unlike the results in Van Dyke and McElree (2011), our results suggest that semantic interference also arises for distractors that do not additionally match the syntactic cue. This finding is not compatible with the proposal that semantic interference only occurs when the distractor additionally shares the syntactic feature with the target noun in (Van Dyke & McElree, 2011).

Looking at the critical regions alone, we observed different patterns in the timing of the effects in English and German. Syntax-first accounts, which assume that syntactic information takes priority over semantic information during processing, could in principle account for the one-word lag between the syntactic and semantic interference effects in German. This includes the proposals that syntactic cues may "gate" semantic cues during retrieval, or that in some configurations, such as antecedent-reflexive dependencies, they may be weighted more highly, or take complete precedence over non-structural cues (Cunnings & Sturt, 2014; Dillon et al., 2013; Kush, 2013; Sturt, 2003; Van Dyke & McElree, 2011). However, these proposals fail to account for the simultaneous effects we observed in English.

One way to reconcile the divergent theoretical positions on the syntax-first debate, and the divergent findings in the critical regions of our data from English versus German, is by assuming a degree of stochastic variability in the use of syntactic and semantic cues. In principle, both types of cues can be used simultaneously, but there could be a lag in the use of semantic cues that occurs probabilistically, leading to delayed semantic effects. However, in the discussion, we foreshadowed a caveat: The different patterns in English versus German must be considered in light of the observed pre-critical effects in both languages.

Thus, a limitation of our findings at the critical verb region is that the effects observed in the pre-critical region, which may or may not be related to retrieval, may have spilled over. To investigate this possibility, we conducted additional exploratory analyses to attempt to statistically control for spillover (Vasishth & Lewis, 2006). Specifically, we reanalyzed regression-path duration at the critical region using the log-transformed and centered RPD at the pre-critical adverb in the same trial as a predictor. For the syntactic manipulation, the resulting estimates were highly similar to those shown in Figure 2.5 (English critical 95% CrI $[-1, 53]$ ms; German post-critical 95% CrI $[1, 39]$ ms). For the semantic manipulation, the posterior means have a positive sign, but the 95% credible intervals include both negative and positive values (English 95% CrI $[-14, 31]$ ms; German 95% CrI $[-9, 26]$ ms). This might suggest that there was indeed some spillover from the pre-critical region, which, however, does not entirely explain away the effects observed at the critical region. Clearly, the syntactic and semantic effects observed here, and in previous work (Van Dyke, 2007), should be replicated with an improved design in a future study. To this end, one possibility would be to further increase the distance between the right clause boundary of the embedded clause(s) and the critical verb, as well as between the verb and the preceding adverb.

A step in this direction was taken by Wagers (2008), who adapted the design of Van Dyke and Lewis (2003) and Van Dyke (2007) to investigate syntactic interference. This design eliminated the potential confound of two- vs one clause boundaries that immediately preceded the critical verb. In addition, the manipulated region included a prepositional phrase across conditions that reduced the possibility of spillover effects from the manipulated distractor regions on the critical regions. This study reported null results at the critical auxiliary and verb. Due to the relatively small sample size of this study, the null result does not resolve the issue (Vasishth et al., 2018).

A general conclusion to be drawn is that when reading patterns at a specific region of the sentence are interpreted, the reading patterns in preceding regions need to be considered as well. A good practice would be to always plot the relevant measures for several regions to the left and right of the critical region, or across the entire sentence (e.g., as in Paape et al., 2021; Paape et al., 2018; Villata et al., 2018). This way, the evolution of processing from one region to the next is made transparent, and potential confounds can be identified.

We now discuss the possible implications of the syntactic and semantic effects that were observed before participants reached the critical verb region.

2.7.1 The possible origins of effects at the pre-critical adverb

In English as well as German, reading time slowdowns due to syntactically and semantically similar distractors were observed at the pre-critical adverb *frequently/*

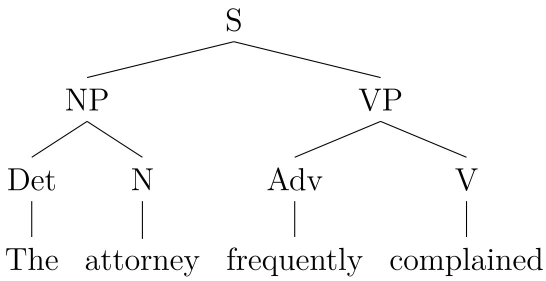
tatsächlich, as shown in Figure 2.5. As mentioned above, Van Dyke (2007) also reported a semantic interference effect at a pre-critical adverbial phrase (Experiment 3). Although in their study, the adverbial was part of an embedded clause. Such pre-critical effects are unexpected under cue-based retrieval accounts, which predict similarity-based interference to arise at the critical verb, where retrieval is assumed to occur.

Recall that the pre-critical adverb in our experiments was added to "absorb" potential clause boundary effects. Therefore, one explanation is that the pre-critical syntactic effects are a consequence of processing two versus one clause boundaries, such that the processing of two clause boundaries led to the reading time slowdown (Arnett & Wagers, 2017; Wagers, 2008).

As both syntactic and semantic effects are observed, there is another explanation for the pre-critical effects: The observed slowdowns caused by matching distractors could be a consequence of faulty encoding of one (or more) noun phrases, that is, by encoding interference (Oberauer & Kliegl, 2006). The memory model of Oberauer and Kliegl (2006) assumes that memory items with similar features—such as the target and distractor nouns in our study—will compete for shared features, which can have detrimental effects on the quality of the representations (Lange & Oberauer, 2005; Nairne, 1990; Neath, 2000). Encoding interference effects have been shown in a number of studies (e.g., Koesterich et al., 2021; Smith et al., 2021; Villata et al., 2018). These effects may manifest during item encoding, or at a subsequent region, such that encoding interference would be a plausible explanation for our pre-critical effects. In fact, recent computational modeling work has independently shown that encoding interference modulates retrieval (Yadav, Smith, et al., 2021).

To our minds, an alternative, plausible explanation is that the early syntactic and semantic effects may be the result of predictive processing. The cue-based retrieval model as implemented by Lewis and Vasishth (2005) specifies a left-corner parsing algorithm that follows X-bar rules (Chomsky, 1986) for incremental syntactic structure building. The left-corner parser operates according to bottom-up and top-down principles. Assuming a context-sensitive phrase structure grammar, once the left corner of the right-hand side (RHS) of a phrase structure rule is identified, the upcoming structure is predicted (e.g., Brasoveanu and Dotlačil, 2020). Specifically, it is assumed that a structure like (6) is recognized as follows:

(6)



Given the NP *The attorney* as the input, the phrase structure rule that has *NP* as the left corner of its RHS is identified ($S \longrightarrow NP\ VP$). The *S*-node of the phrase-structure grammar is then predicted top-down. Then, given the input of the next word, *frequently*, the rule that has *Adv* as the left corner of the RHS is identified bottom-up ($VP \longrightarrow Adv\ V$), and *VP* is predicted. As the *Adv* has been integrated into the structure and *VP* is already predicted, it is possible that the retrieval of a subject is already triggered at *frequently*. This is because, at the predicted VP, a grammatical subject is required, and a {*subject*} cue could initiate a search for a +*subject* feature.

The problematic aspect of this proposal is that the adverb would need to set retrieval cues when the identity of the verb is still unknown. It is thus unclear why a semantic feature like {*animate*} should be used in the retrieval, as the upcoming verb could, in theory, also accept inanimate subjects. As an additional assumption, one could hypothesize that the processor has a probabilistic expectation for an upcoming verb with an animacy cue, given that subjects are frequently animate (Clark & Begun, 1971; Kathryn Bock & Warren, 1985), and thus the cue is put to use independently of the identity of the verb. Another possibility is that participants get attuned to expecting verbs with an animacy cue through repeated exposure to the items. Both possibilities could be in operation in tandem, leading to retrievals being triggered in the pre-critical region.

Pre-verbal structure building has been shown in a number of verb-final languages (e.g., Aoshima et al., 2004; Bader and Lasser, 1994; Kamide and Mitchell, 1999; Konieczny, 2000), as well as English, a verb-medial language (Omaki et al., 2015). For example, Omaki et al. (2015) found that in filler-gap dependency resolution, the parsing mechanism actively makes a prediction of the gap position prior to accessing the verb properties. These findings are compatible with a view that a subject-retrieval may occur at a pre-verbal modifier.

Findings by Wagers and McElree (2009) also seem to be in line with the patterns we observe on the pre-critical adverb. In two SAT experiments, the authors showed that in sentences such as (7), the presence of VP-level adverbs resulted in a processing speedup on the verb, compared to when no modifier was present.

(7) The officer was informed that the driver (abruptly) fainted.

The speedup was observed for VP-level adverbs, but not for S-level adverbs (evaluative- or epistemic modality adverbials). Wagers and McElree proposed that the speedup occurs because verb processing is given a "head start" by VP-level adverbs. Their "head-start" hypothesis would be compatible with our hypothesis that subject-retrieval may be initiated at the adverb, although our items contain both VP- and S-level adverbs.

Although this explanation seems plausible, our experiments were not designed to test interference effects at verb modifiers. Therefore, further investigation is necessary to test if these effects are reproducible, and to explore the precise conditions under which they arise.

Overall, our data from two languages provide additional support for syntactic and semantic similarity-based interference in online processing. The findings are compatible with the hypothesis that syntactic and semantic interference effects can arise simultaneously in online sentence comprehension. Surprisingly, the effects arose prior to the critical verb in both languages. These results are compatible with cue-based theories of parsing, given the assumption that retrievals can occur on pre-verbal modifiers. In addition, our offline data also show patterns consistent with semantic interference, indicating that semantically similar distractors can interfere with overall comprehension.

2.8 Conclusion

Our cross-linguistic investigation in English and German presents support for syntactic and semantic interference in sentence comprehension. The data suggest that syntactic and semantic interference can arise simultaneously during the formation of long-distance dependencies. These findings are compatible with models that assume the simultaneous use of syntactic and semantic information. Additionally, in both languages, our offline comprehension question data suggest that semantic interference can adversely affect overall comprehension.

Appendix A

Question responses

Recall that a third of questions in our experiments targeted NP1 (the target, e.g., *the attorney*), one third targeted NP2 (distractor, e.g., *the secretary*), and another third targeted NP3 (the manipulated distractor, e.g., *visitor/meeting*). Interestingly, when the target NP1 (e.g., *the attorney*) was the correct response, participants frequently erroneously chose NP2 (e.g., *the secretary*), and vice versa. This tendency seems to be stronger in English than in German (Tables A.1 and A.2).

English:				
A) Responses in per cent (%) to questions targeting NP1 (the attorney)				
Condition	**NP1** (attorney) **correct**	NP2 (secretary) incorrect	NP3 (visitor) incorrect	'I don't know' (?) incorrect
a. −subject, −animate	60	39	0	1
b. −subject, +animate	53	38	7	2
c. +subject, −animate	58	39	0	3
d. +subject, +animate	49	35	12	4
B) Responses in per cent (%) to questions targeting NP2 (the secretary)				
Condition	NP1 (attorney) incorrect	**NP2** (secretary) **correct**	NP3 (visitor) incorrect	'I don't know' (?) incorrect
a. −subject, −animate	27	71	0	2
b. −subject, +animate	26	59	9	6
c. +subject, −animate	29	70	0	1
d. +subject, +animate	23	64	10	3
C) Responses in per cent (%) to questions targeting NP3 (the visitor/meeting)				
Condition	NP1 (attorney) incorrect	NP2 (secretary) incorrect	**NP3** (meeting/ visitor) **correct**	'I don't know' (?) incorrect
a. −subject, −animate	10	8	76	6
b. −subject, +animate	12	10	68	10
c. +subject, −animate	10	7	80	3
d. +subject, +animate	17	13	66	4

Table A.1: English: Given responses (in %) by condition for questions that targeted A) NP1 (*the attorney*), B) NP2 (*the secretary*), or C) (*the meeting/visitor*).

German:

A) Responses in per cent (%) to questions targeting NP1 (the journalist)				
Condition	**NP1** (journalist) **correct**	NP2 (colleague) incorrect	NP3 (mafia boss) incorrect	'I don't know' (?) incorrect
a. −subject, −animate	84	14	0	2
b. −subject, +animate	76	15	6	3
c. +subject, −animate	86	13	0	1
d. +subject, +animate	77	11	7	5
B) Responses in per cent (%) to questions targeting NP2 (the colleague)				
Condition	NP1 (journalist) incorrect	**NP2** (colleague) **correct**	NP3 (mafia boss) incorrect	'I don't know' (?) incorrect
a. −subject, −animate	22	69	0	9
b. −subject, +animate	25	56	9	10
c. +subject, −animate	18	75	0	7
d. +subject, +animate	17	61	9	13
C) Responses in per cent (%) to questions targeting NP3 (the scandal/mafia boss)				
Condition	NP1 (journalist) incorrect	NP2 (colleague) incorrect	**NP3** (meeting/ mafia boss) **correct**	'I don't know' (?) incorrect
a. −subject, −animate	13	9	69	9
b. −subject, +animate	17	9	61	13
c. +subject, −animate	12	4	75	9
d. +subject, +animate	21	5	63	11

Table A.2: German: Given responses (in %) by condition for questions that targeted A) NP1 (*the journalist*), B) NP2 (*the colleague*), or C) (*the scandal/mafia boss*).

Chapter 3

Is there cross-linguistic evidence for proactive cue-based retrieval interference in sentence comprehension? Eye-tracking data from English, German and Russian

Daniela Mertzen, Anna Laurinavichyute, Brian W. Dillon, Ralf Engbert, and Shravan Vasishth (2021b). Unpublished manuscript.

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 the verb. When these retrieval cues match not only a target item, but also similar items in memory, this leads to similarity-based interference. In the majority of studies on interference, the effects may result from cue-dependent retrieval, but may also be explained by erroneous item encodings. A landmark study by Van Dyke and McElree (2006) explicitly tested cue-based retrieval interference: retrieval cues were manipulated to match or mismatch semantically similar items that were presented proactively (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 large-scale cross-linguistic investigation of proactive cue-based retrieval interference. We present three larger-sample eye-tracking studies in English, German, and Russian that tested cue-based interference in the online processing of filler-gap dependencies under varying task demands. The English study finds very weak support for proactive cue-based interference in online dependency resolution, but only under low task demands. By contrast, the German and Russian data showed some evidence against the interference effect, suggesting that in languages with richer case marking, interference might be attenuated. Overall, there is some very modest support in our English data for proactive interference that is cue-based and has its source at retrieval. Language-specific differences are not explicitly predicted by cue-based accounts and need further investigation.

3.1 Introduction

Sentence comprehension requires us to rapidly form dependencies between non-adjacent words. For example, in (8) a dependency needs to be established between the verb *complained* and its subject *the resident* (Van Dyke, 2007).

(8) *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 influential account, the cue-based retrieval account, 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 (8), 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.

A consequence of this cue-based retrieval process is similarity-based interference: the retrieval of a target item may be impeded by other items in the sentence, called distractors, which carry similar syntactic or semantic features as the target.

In (8), *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 encoding in working memory, which creates similarity-based interference. These interference effects are reflected in increased processing times in reading studies.

There is considerable evidence for similarity-based interference in sentence comprehension. Many reading studies have investigated interference from distractors that intervene between the retrieval target and the retrieval site, as in example (8) (*retroactive* interference; 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). Other studies report interference from within-sentence distractors that precede the critical dependency (*proactive* interference; e.g., Cunnings & Sturt, 2014; Koesterich et al., 2021; Parker & Phillips, 2017; Van Dyke & McElree, 2011; Wagers et al., 2009). For both pro- 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). Although it has been suggested that proactive interference may affect dependency formation to a lesser degree than retroactive interference (Jäger et al., 2017; Van Dyke & McElree, 2011).

In the pre- and retroactive interference literature, many studies have reported that words encoded in memory can interfere with the processing of linguistic dependencies if they are semantically similar (e.g., Cunnings & Sturt, 2018; Glaser et al., 2013; Gordon et al., 2001, 2004; Gordon et al., 2006; Laurinavichyute & von der Malsburg, 2021b; Lowder & Gordon, 2014; Rich & Wagers, 2020; Van Dyke, 2007). Here, we focus on one line of research that investigates a special case of proactive semantic interference: interference from sentence-external items that were encoded in memory 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 show that sentence parsing can even be disrupted by linguistic items that are not structurally integrated. This research addresses a broad theoretical question in the sentence processing literature, namely, whether sentence parsing relies on domain-general or domain-specific memory resources (Caplan & Waters, 1999; Just & Carpenter, 1992). Interference from sentence-external items on real-time sentence comprehension supports a view that sentence parsing relies on general memory resources used for other cognitive tasks. While these studies have shown some support for similarity-based interference effects, it remains unclear whether there is convincing evidence for interference that unambiguously arises due to cue-overload during retrieval, i.e., *cue-dependent* retrieval interference.

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. 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, processing times were increased.[1]

Fedorenko et al. (2006) investigated the Gordon et al. (2002) claim in non-clefted subject- and object-extracted relative clause structures. Their design used a memory load of either one or three nouns. Fedorenko and colleagues reported that when three nouns had to be memorized, interference was greater for object than subject-extracted relative clauses.

While the design of the Gordon et al. (2002) and the Fedorenko et al. (2006) studies can test for similarity-based interference effects, the design does not allow us to determine the source of the interference effects. Interference effects that

[1]While this main effect of match reached statistical significance, the critical interaction between match and cleft type did not; Gordon et al. (2002), p. 429.

are observed with this study design would be compatible with cue-based retrieval accounts, as well as encoding accounts (Lewandowsky et al., 2008; Oberauer & Kliegl, 2006). Encoding accounts postulate that it is not cue overload at the retrieval site but the erroneous encoding of similar linguistic items in memory that leads to interference. These erroneous encodings can result from feature loss (Oberauer & Kliegl, 2006). If deteriorated representations of NPs in memory lead to interference, then these encoding interference effects may also be observable at the retrieval site.

A landmark study by Van Dyke and McElree (2006) adapted the Gordon et al. (2002) study design, keeping the sentence-external distractors constant, and instead manipulating the retrieval cues at the verb (the point of retrieval). This design can explicitly test *cue-dependent retrieval interference* from semantically-similar, sentence-external distractors. Table 3.1 shows the Van Dyke and McElree (2006) design: 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*})[2] 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 a 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 includes baseline comparison 'No memory load' conditions, showing the same sentences as the Memory load conditions, without presenting the memory nouns. In No load conditions, no significant reading time differences are expected. This predicted reading time pattern in Load and No load conditions would be reflected in an interaction between the factors *Memory load* and *Interference*.

Only cue-based retrieval accounts predict this interaction. By comparison, encod-

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

ing accounts predict no reading time differences in the Memory load conditions because retrieval cues do not play a role for encoding interference. As the same target NP and memory words are used across conditions, degraded memory encodings for these items should lead to the same amount of interference in the two conditions, i.e., no difference in reading times would be expected. It is important to note that for this design, a null result does not allow us to draw conclusions about the presence or absence of encoding interference. A null result can only tell us that there is no support for retrieval interference. Therefore, this manipulation would show clear support for interference that has its source at retrieval: 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 3.1: 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 <u>*sailed*</u> in two sunny days.

Interference
It was *the boat* that the guy who lived by the sea <u>*fixed*</u> in two sunny days.

Van Dyke and McElree (2006) reported a reading time pattern consistent with cue-based retrieval interference from extra-sentential items on within-sentence dependency resolution. However, the evidence for this effect was rather weak (significant by participants and items: $F_1(1, 55) = 4.07$, $p < 0.04$, $F_2(1, 35) = 5.58$, $p < 0.02$; non-significant $minF'$ statistic: $minF'(1, 90) = 2.35$, $p = 0.13$).[3] A subsequent study with the same design and stimuli did not show any evidence for this interference effect in online sentence comprehension (Van Dyke et al., 2014). Thus, there is inconclusive evidence for cue-dependent retrieval interference from sentence-external materials in online sentence comprehension. To date, the claim for this effect rests on self-paced reading data in English.

Overall, while there is evidence for semantic interference effects in online sen-

[3]Significance of $minF'$ is required to show that an effect can be generalized over participants as well as items (Clark, 1973).

tence comprehension, it remains unclear whether those effects are driven by cue-dependent retrieval processes. The present study is motivated by this current state of evidence: we investigated proactive cue-dependent retrieval interference in filler-gap dependencies, and aimed to extend findings across languages and tasks.

3.1.1 Motivation for the present study

The goal of this study was to test proactive cue-based retrieval interference from extra-sentential material on intra-sentential dependency resolution (henceforth ‘proactive interference’), using a dual-task paradigm (Van Dyke & McElree, 2006). This is important because the Van Dyke and McElree (2006) manipulation explicitly tested for cue-based retrieval interference, excluding encoding interference as a possible alternative explanation. Studying interference effects in such a paradigm is thus key in determining the role that cue-dependent retrieval interference plays in the processing of linguistic dependencies. Support for cue-overload due to sentence-external distractors would strengthen cue-based accounts which posit that a cue-based retrieval mechanism guides dependency formation during real-time sentence comprehension. Specifically, this effect would point to a retrieval mechanism that is not constrained to the sentence context, but that erroneously considers extraneous linguistic items as retrieval targets.

3.1.1.1 Re-investigation of proactive interference

The reason for the inconclusive findings in Van Dyke and McElree (2006) and Van Dyke et al. (2014) could be that the phenomenon under investigation exhibits only subtle effects. This is likely, given the results of Jäger et al. (2017) who derived quantitative estimates by conducting a meta-analysis of published studies on similarity-based interference effects in sentence comprehension. For non-agreement subject–verb dependencies, they showed an interference effect that lies between 2 and 28 ms.[4] Small effects like similarity-based interference can remain undetected

[4] All 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) entered the meta-analysis for non-agreement subject-verb dependencies. Of those studies, only the Van Dyke and McElree (2006) study used a dual-task paradigm with sentence-external

if tested with participant sample sizes routinely used in experimental psycholinguistics (e.g., Jäger et al., 2017; Vasishth et al., 2018). Prospective statistical power (i.e., the probability of detecting an effect of a certain magnitude in a planned experiment) is low for interference effects with standard participant sample sizes of 40-60 subjects (Jäger et al., 2017). Figure 3.1 shows that even with a participant sample size of 120, one can end up in a low power scenario when investigating small effects.

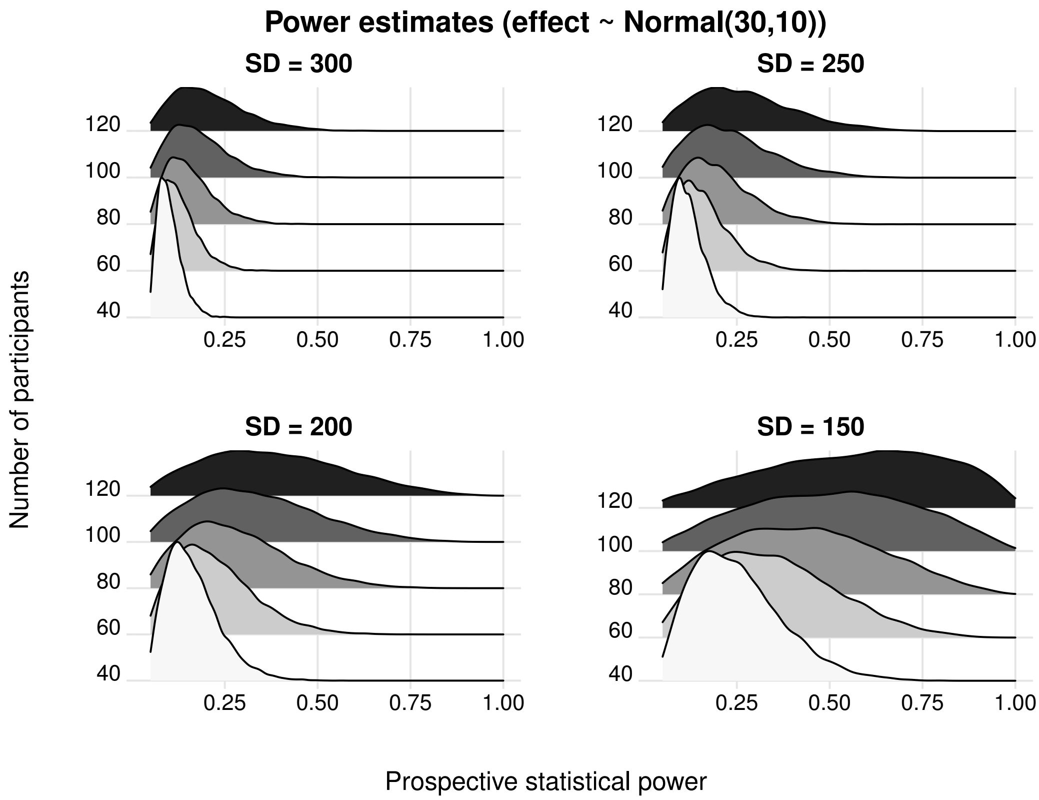


Figure 3.1: Power estimates for an effect size $\mathcal{N}(30, 10)$ assuming various residual standard deviations (SD) (300, 250, 200, 150) for a different number of participants (40, 60, 80, 100, 120). Different shadings are merely for better distinguishability. The figure is adapted from Vasishth (2019).

Low power studies lead to significant effects that are greatly overestimated (Type M (Magnitude) error; Gelman and Carlin, 2014). The interference effect reported in the Van Dyke and McElree (2006) study with sample size 56 could be such an inflated estimate (for the critical interaction: 41 ms with SE 20 ms). There

distractors. The attentional demands of this task are quite different from a common reading task. However, all of these studies investigated non-agreement subject-verb dependencies, and the effect magnitudes from these studies are relatively comparable.

are several large-sample studies in the psycholinguistic literature that suggest that inflated estimates from underpowered studies may not be replicable. These studies re-investigated various psycholinguistic phenomena and were not able to replicate the original effects (e.g., Nicenboim et al., 2020; Nieuwland et al., 2018; Vasishth et al., 2018). The current study also aimed for a large participant sample size to obtain a precise estimate of the proactive interference effect.

3.1.1.2 A cross-linguistic investigation of proactive interference

Similarity-based interference in sentence comprehension has 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, the majority of research has focused on English, including the work on proactive interference from extra-sentential distractors (e.g., Cunnings & Felser, 2013; Cunnings & Sturt, 2014, 2018; Dillon et al., 2013; Sturt, 2003; Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006, 2011). If cue-based parsing mechanisms are an integral part of human language processing, as posited by cue-based theories, then proactive interference should be observable not only in English but cross-linguistically.

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).

Several findings in the literature are consistent with the hypothesis that overt case marking reduces retrieval interference. Much of the support for this hypothesis comes from language production studies that investigate so-called agreement attraction errors. For example, given a complex NP such as *The key to the cabinets* frequently results in the production of a plural verb given the plural "attractor" NP *the cabinets*. In a production study on this phenomenon in English, participants were asked to produce verbs for complex NPs that were presented to them (Nicol & Antón-Méndez, 2009). The authors reported that fewer attraction errors were produced for complex NPs with a singular head that had a prepositional phrase (PP) modifier which was marked for accusative case (such as *the letter from him/them*), compared to a PP modifier that did not have overt case marking (such as *the letter from the editor(s)*). A study on German also showed a pattern suggestive of reduced attraction for overt and unambiguous case marking (Hartsuiker et al., 2003). Hartsuiker et al. (2003) manipulated number (singular, plural) and case marking (Nominative-Accusative ambiguous, Dative unambiguous) on the PP modifier of a complex NP. This study reported that attraction errors were reduced in the unambiguous conditions (German: *Die Stellungnahme zu der/den*$_{DAT}$ *Demonstration(en)*, English: *The position on the demonstration(s)*) compared to the ambiguous conditions (German: *Die Stellungnahme gegen die*$_{NOM/ACC}$ *Demonstration(en)*, English: *The position against the demonstration(s)*).

However, the results from a recent study on number agreement attraction effects in Eastern Armenian suggests that case may not play a role (Avetisyan et al., 2020). This study found no indication that case marking interacts with number attraction, although a main effect of case was reported where faster reading times were observed for distinct case marking in both grammatical and ungrammatical sentences. If, in our study, case marking does not play a role, then proactive interference effects should be of the same magnitude across languages.

3.1.1.3 Proactive interference under different processing depths

The similarity-based interference prediction is contingent on a simple but important assumption that is implicit in cue-based retrieval theories: That all syntactic dependencies are resolved during real-time processing. However, another sentence processing theory, good-enough processing, assumes that syntactic dependencies are not always resolved, but can remain underspecified (Ferreira et al., 2002). 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, 2016).

Swets et al. (2008), for example, implemented an experimental design by Traxler et al. (1998), using sentences involving ambiguous relative clause attachments to study the effects of varying number and complexity of questions on (self-paced) reading times, testing globally and locally ambiguous sentences as in Example 9.

(9) a. The maid of the princess who scratched herself . . . was terribly humiliated. (global)

b. The son of the princess who scratched himself . . . was terribly humiliated. (high)

c. The son of the princess who scratched herself . . . was terribly humiliated. (low)

Their between-subjects question manipulation consisted of three question-conditions: (1) complex comprehension questions targeting the readers' interpretation of the relative clauses in each trial, (2) superficial comprehension questions targeting a non-critical region in each trial, or (3) superficial questions, but only for a subset of trials. In the simple-question conditions, Swets et al. (2008) reported a so-called 'ambiguity advantage', that is, faster reading times for ambiguous compared to unambiguous sentences (Traxler et al., 1998; Van Gompel et al., 2005; Van Gompel et al., 2001). Specifically, globally ambiguous sentences were read faster than both locally ambiguous high- and low-attachment sentences. In the complex-question conditions, this advantage for globally ambiguous sentences was reduced. Here, both globally ambiguous and locally ambiguous low-attachment sentences were read faster than high-attachment sentences. Swets et al. (2008)

concluded that the commitment to attach or not is in part dependent on the goal the reader pursues.

If comprehenders underspecify syntactic relations given a simple task, the finding by Swets et al. (2008) on relative clause attachment ambiguities may generalize to other linguistic dependencies. We build on the finding of task-dependent performance by testing the depth-of-processing hypothesis using a within-subjects manipulation for filler-gap dependencies. If superficial processing can be induced by a simple task, then it would be expected that proactive interference effects are smaller in magnitude under 'superficial' compared to 'deep' processing conditions, or the effects may disappear altogether.

3.2 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 3.2.

3.2.1 Experimental design and materials

3.2.1.1 The Memory load by 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 our study is that Van Dyke and McElree (2006) tested object-cleft sentences whereas our study

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 3.2: 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.

uses non-clefted structures with two embedded relative clauses. In our materials, the target object is not in linguistic focus, whereas the clefted object retrieval targets in Van Dyke and McElree (2006) are. 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). This may reduce the magnitude of an interference effect. Thus, we expected that non-clefted stimuli may be advantageous to detect the predicted effect.

Table 3.3 shows English example items for the complex (3.3A) and the simple (3.3B) 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. It has been reported that multiple center-embedded sentences may be read like a list of unrelated words which may reduce comprehension (e.g., see Miller, 1962; Miller and Chomsky, 1963). To avoid 'list-like' sentences (such as 'NP1 NP2 NP3 VP3 VP2 VP1'), and to increase the naturalness of the sentences, a prepositional- or an adverbial phrase was added before the matrix-clause verb.

In (3.3A), the critical dependency is between the relative clause verb *sailed/fixed*

Table 3.3: English example items (complex and simple version).

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 ...

Interference

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

No memory load: — — —

No interference

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

Interference

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

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

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). For the complex experiment version, we used 40 experimental items, and 90 filler items. For the simple experiment version, we created 40 new experimental items and 90 new filler items. Examples of the German and Russian experimental items are shown in Appendix B; a description of filler sentences can be found in Appendix C.

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*).

Plausibility was rated for all 80 items in each language, 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 10. 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/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).

(10) 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 3.2 shows the estimated probabilities for each of the seven rating choices in each condition. To estimate the probabilities, we fit Bayesian cumulative models in brms (Bürkner, 2017; Bürkner and Vuorre, 2019; see supplementary materials). Overall, the pattern is as expected: each language shows high plausibility ratings for a–e, that is, for the verb–target combinations (*fixed/sailed the boat*), and for the verb–plausible distractor combinations (*fixed the table/sink/truck*). By contrast, lower plausibility ratings are observed in f–h, that is, for the verb–implausible distractor combination (*sailed the table/sink/truck*).

3.2.1.2 The Depth of Processing manipulation

Depth of processing was manipulated through comprehension question complexity. In one experiment version, the comprehension 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 sought to induce 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) (e.g., *'Did the guy live by the sea?'*, Table 3.3A). The other half of the questions targeted the main clause subject-verb dependency (e.g., *'Was it the boat that was old?'*).

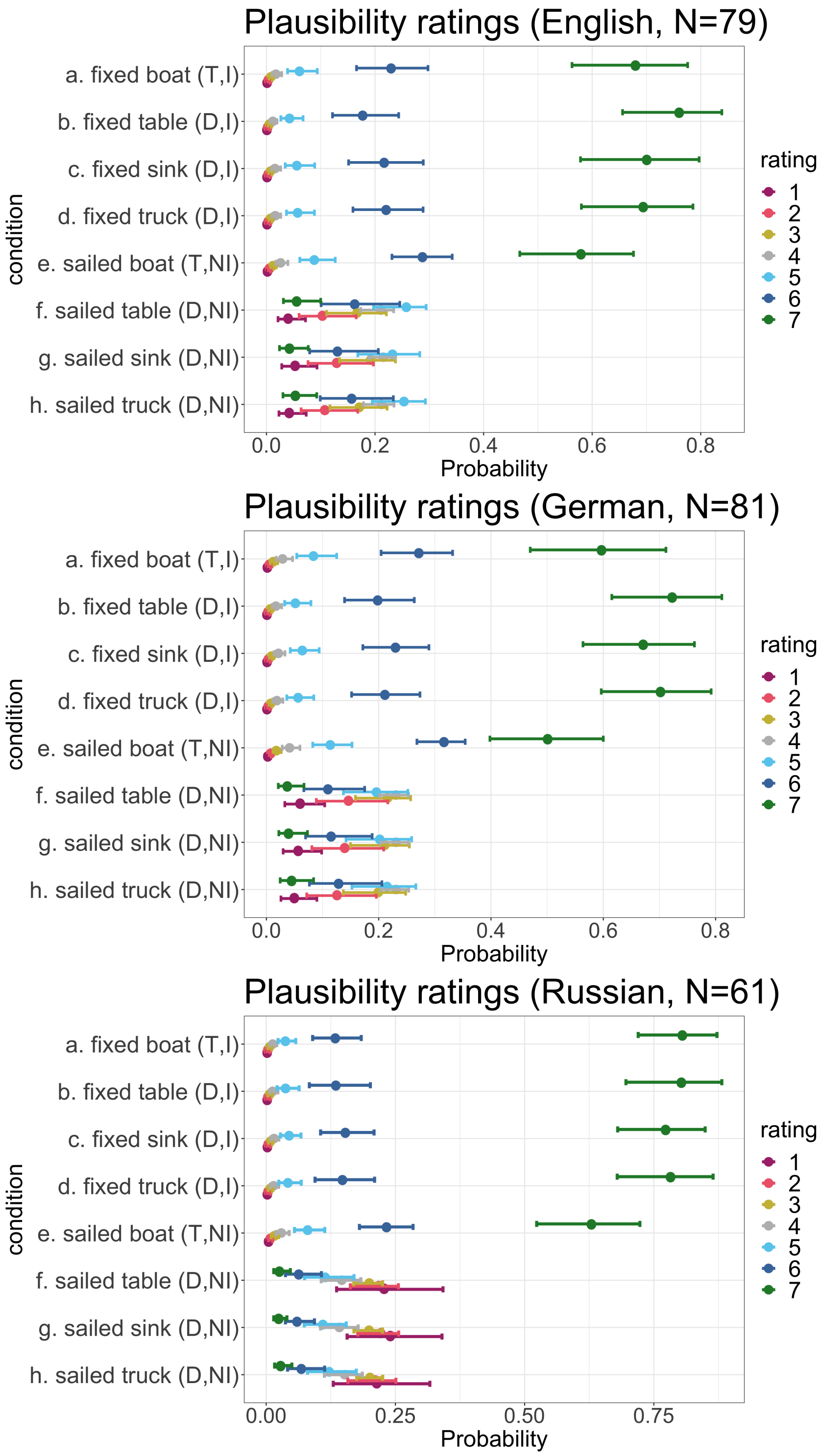


Figure 3.2: Plausibility ratings results for the items used in English, German, and Russian. Shown are the estimates (posterior means with 95% credible intervals) of the probability of the plausibility rating in each condition. '1' = very implausible, '7' = very plausible. T = target, D = distractor, I = Interference, NI = No interference. For ease of interpretability, we use the English condition labels for all three languages.

Simple version In this version, simple questions aimed to induce superficial processing because they did not require participants to resolve the dependencies in the sentence. These questions targeted non-critical NPs in the sentence (e.g., *'Was a pilot mentioned in this sentence?'*, Table 3.3B).

3.2.1.3 English vs. German and Russian materials

Figure 3.3 shows schematic example sentences with memory nouns in the Load conditions 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, this should result in an overloaded semantic cue.

In the English sentences, the target NP and the complementizer *that* are not overtly marked for case. By contrast, in German and Russian, the target NP and relative pronoun have overt case marking. 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.

3.2.1.4 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.

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 ('stopping rule', Kruschke, 2015). For English,

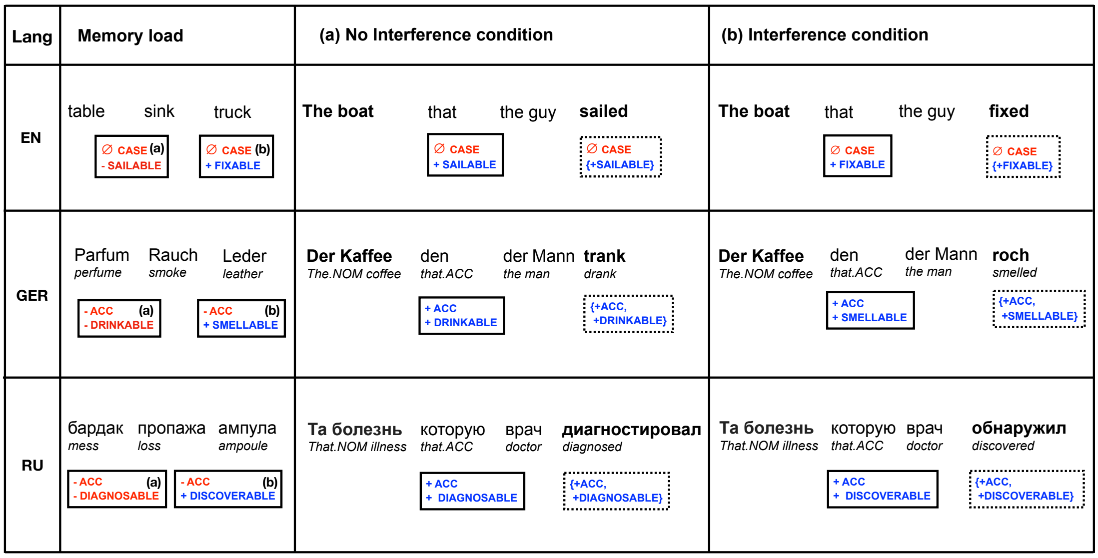


Figure 3.3: Example stimuli (schematic) for the English, German, and Russian experiment.

due to logistical reasons, data collection stopped when we reached a 40 ms interval. For both German and Russian we obtained more precise estimates of the effect, because we were able to test a much larger participant sample. A full overview of the participant profiles is shown in Table 3.4.

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 loss. We excluded individual trials where the fixation sequence was disturbed by external influences. In addition, a small subset of participants was not able to participate in the second part of the study, such that, for those participants, we only have the data of either the complex or the simple experiment version.

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 3.4: 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.

3.2.1.5 Procedure

Participants saw both experiment versions (complex or simple) across 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 close attention should be paid to both the reading and the recall task. 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[5] 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.[6] 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

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

[6] 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 to silent reading was necessary as, for the eye-tracking method even small head movements should be avoided, and head movements were restricted by the headrest.

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. Table 3.5 shows the differences in technical specifications for the English, German, and Russian study.

Study	Location of data collection	Eye-tracker (mount)	Monitor resolution	Eye - screen distance	Font (size)	Chars per degree of visual angle
English	UMass Amherst MA, USA	EyeLink 1000 (tower)	1440×900	64 cm	consolas (16)	4.5
German	Potsdam University, Germany	EyeLink 1000 Plus (tower)	1920×1080	56 cm	consolas (18)	2.4
Russian	Higher School of Economics Moscow, Russia	EyeLink 1000 Plus (desktop)	1920×1080	90 cm	consolas (20)	3.9

Table 3.5: Shown are the location of the eye-tracking lab used for data collection for each study (English, German, Russian), and differences in the technical specifications of the in eye-tracking setup in the different labs. Chars = characters.

3.2.2 Predictions

For all three languages, an interference effect was predicted to occur at the outer RC verb (e.g., *fixed/sailed* in Table 3.3). 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.

3.2.2.1 Predictions for the Depth of processing manipulation (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.

3.2.2.2 Predictions for the cross-linguistic comparison between German and Russian vs. English

In German and Russian, the Load × Interference interaction was expected to be smaller than in English. The reason for this is 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.

3.2.3 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 × 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., 2017). Memory load, Interference and their interaction were included in the models as fixed effects. The models had full variance-covariance matrices for subjects and for items 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 3.6. 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 Normal(0,10). All other parameters were defined as a standard Normal(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 values; this is illustrated in Figure 3.4. A ν parameter value > 1 will favor less correlation between intercepts and slopes, which seems reasonable given that we do not have any information about the correlations. For each of the statistical models, we ran four chains, each with 4000 iterations. The first half of these samples were discarded as warm-up, or burn-in, samples. The $\hat{R}$-diagnostic (Gelman et al., 2014) as well as visual inspection of trace plots provided us with information about model convergence.

We report the 95% Bayesian credible intervals (CrI) of the posterior distributions. These tell us the range within which the true parameter lies with a 95% probability, given the data and model. For easier interpretation the values were back-transformed from the log scale to the millisecond scale. For the fixed effects in the model (i.e., $\log(rt) = \beta_0 + \beta_1 Load + \beta_2 Interference + \beta_3 Interaction$) with

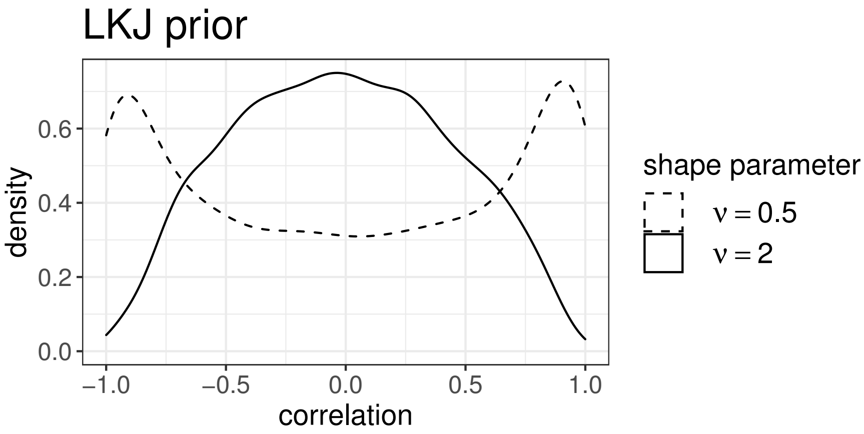


Figure 3.4: The dashed line shows that when the ν parameter of Stan's LKJ function is 0.5, less extreme correlation values are downweighted. The solid line shows that changing the ν parameter value to 2.0 downweights more extreme values.

contrast coding $\pm.5$, the posterior of the difference in factor level means on the log scale was back-transformed in the following way: $\exp(\beta_0 + .5 \times \beta_1) - \exp(\beta_0 - .5 \times \beta_1)$ (Nicenboim et al., 2021).

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

Table 3.6: Contrast coding for effects of Load, Interference, and their interaction.

3.2.3.1 Bayes factor analyses for model comparison

We conducted Bayes factor analyses to evaluate the evidence in favor of one model over another model (Gelman et al., 2014; Jeffreys, 1961; Kass & Raftery, 1995; Lee & Wagenmakers, 2014; Schad et al., 2021). Here, 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). 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:

$$BF = \frac{P(Data|Model_0)}{P(Data|Model_1)} \tag{3.1}$$

The Bayes factor can 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. Jeffreys (1961) as cited in Lee and Wagenmakers (2014) gives a guideline for the interpretation of the Bayes factor (Table 3.7, minimally adapted from Lee and Wagenmakers, 2014, p. 105).

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 3.7: Guideline 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.

The Bayes factor is sensitive to the prior distribution (Gelman et al., 2017; Sinharay & Stern, 2002). Mildly informative priors such as $\mathcal{N}(0, 1)$ very likely bias the Bayes factor in support of a model that reflects the hypothesis that the parameter of interest is zero (Mulder & Wagenmakers, 2016; Rouder et al., 2018). We therefore computed the Bayes factor for more informative priors on the interaction term in Model 1, $\mathcal{N}(0, 0.1)$ and $\mathcal{N}(0, 0.05)$ (see also Nicenboim et al., 2020).

The models were fit using the R package `brms`, an interface using Stan to fit Bayesian hierarchical models (Bürkner, 2017). For each model, we ran four chains with 10000-30000 iterations each. The first 2000 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, Sarafoglou, et al., 2017; Gronau, Singmann, et al., 2017).

3.2.4 Results

3.2.4.1 Memory recall

Figure 3.5 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 3.5A are based on a strict criterion where recall was judged as 'correct' only when all three memory nouns were recalled in the correct order.[7] 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 3.5B. The lower recall accuracy in our study may be the result of the participants silently rehearsing 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).

[7] 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; see supplementary materials.)

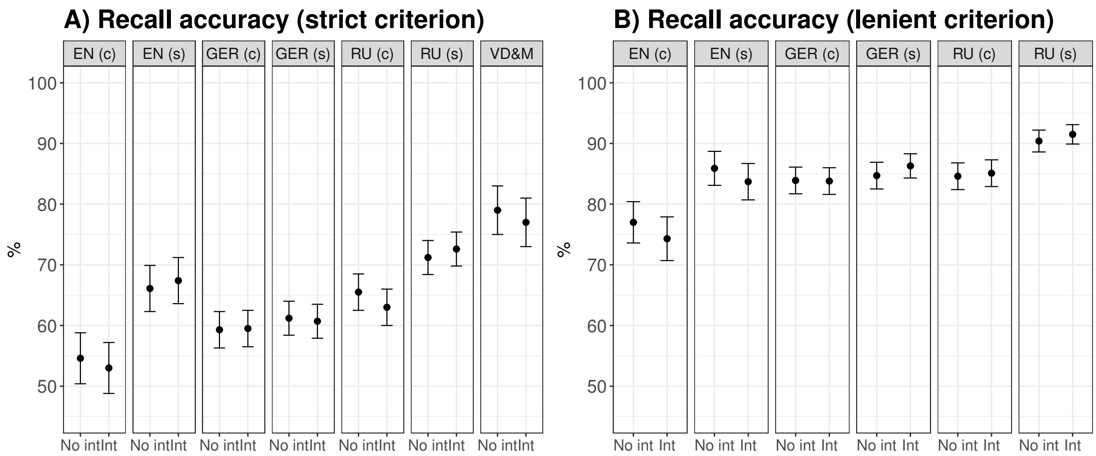


Figure 3.5: Mean recall accuracies (in percent) with 95% confidence intervals, A) for a strict criterion (all three words recalled in the correct order), and B) for a lenient criterion (two or three words recalled in any order). EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version; VD&M = Van Dyke & McElree (2006); No int = No interference condition, Int = Interference condition.

3.2.4.2 Comprehension question accuracy

The by-condition question response accuracies for all six experiment versions are shown in Figure 3.6. Question response accuracies were high across all experiment versions. Filler item accuracies were also high (mean accuracies and standard errors (SE) in English (complex): 91% (SE .5), English (simple): 93% (SE 1), German (complex): 88% (SE .5), German (simple): 97% (SE .2), Russian (complex): 87% (SE .5), Russian (simple): 86% (SE .5).

3.2.4.3 Reading times

We report the results at the critical verb for the reading measures first-pass reading times (FPRT) and total fixation time (TFT).[8] FPRT, also 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 3.7.

[8]We conducted further statistical tests of other commonly analyzed eye-tracking measures, regression-path duration and proportion of first-pass regressions. These are also not reported here as they do not change the conclusions of this paper.

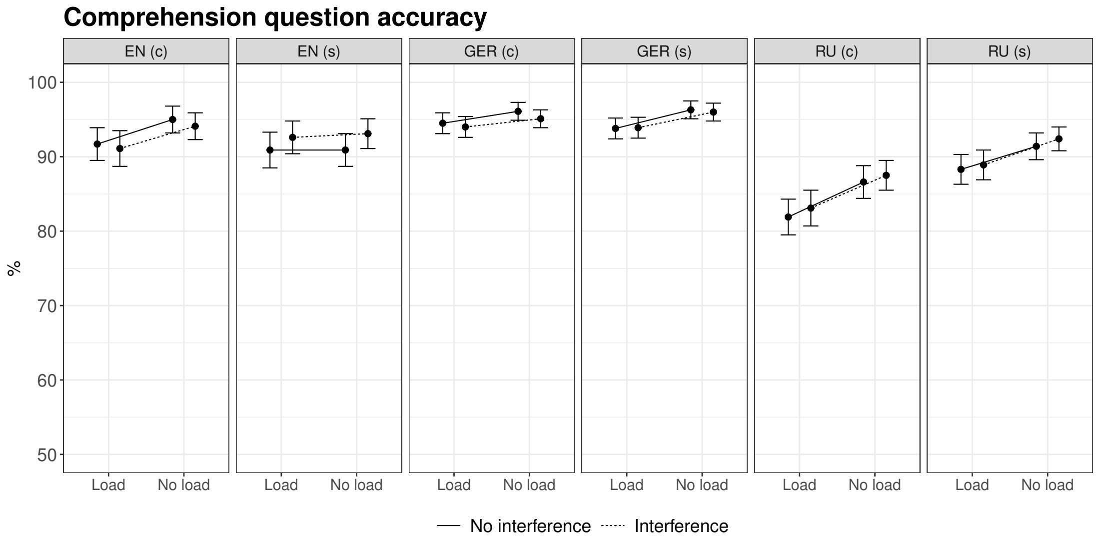


Figure 3.6: By-condition mean comprehension question accuracies (in percent) with 95% confidence intervals. EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version.

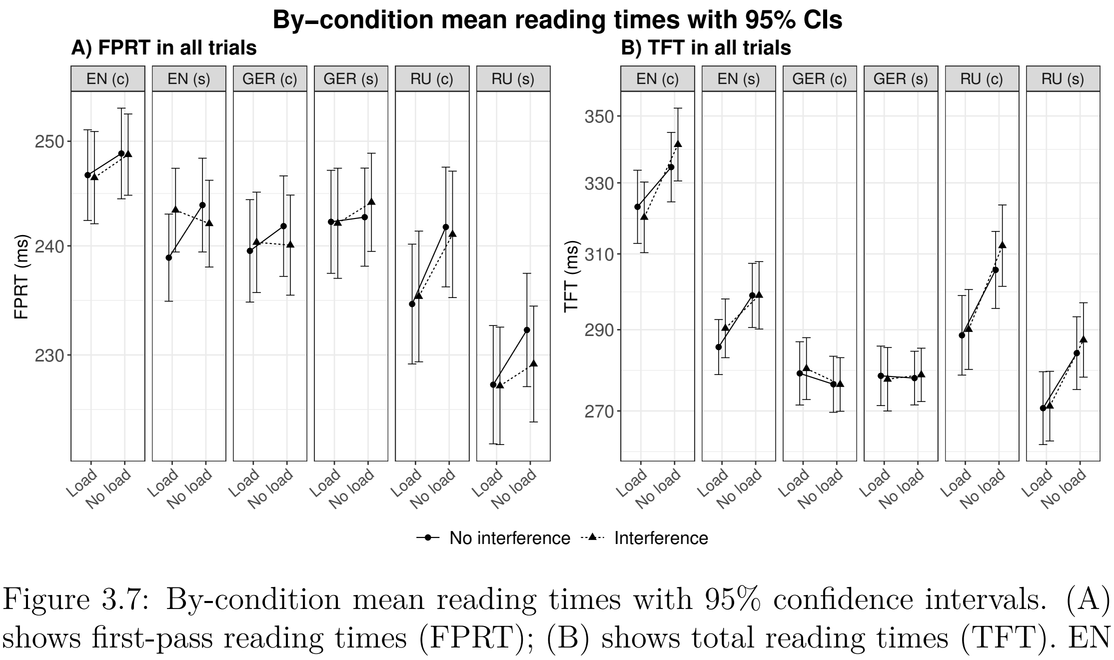


Figure 3.7: By-condition mean reading times with 95% confidence intervals. (A) shows first-pass reading times (FPRT); (B) shows total reading times (TFT). EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version.

Figure 3.8 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, i.e., Load conditions were read faster than No load conditions. 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, 18] 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 [−3, 13] ms).

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 [−11, 1] ms, nested Load: [−15, 12] ms, nested No load: [−4, 21] ms). For TFT, the 95% CrI is [−16, 0] ms (nested Load: [−24, 9] ms, nested No load: [−10, 29] ms). Similarly, Russian (complex) also shows a negative sign for the interaction (FPRT: 95% CrI [−17, −1] ms; nested Load: [−29, 0] ms, nested No load: [−10, 18] ms; TFT: 95% CrI [−27, 1] ms; nested Load: [−20, 29] ms, nested No Load: [0, 63] ms). For all other experiment versions, the Load × Interference interaction is centered around zero.

To summarize, Figure 3.9 shows our FPRT estimates of the interaction from each experiment version next to the Van Dyke and McElree (2006) interaction estimate. Only the simple experiment versions of English and Russian show effect estimates that overlap with the interaction in Van Dyke and McElree (2006).

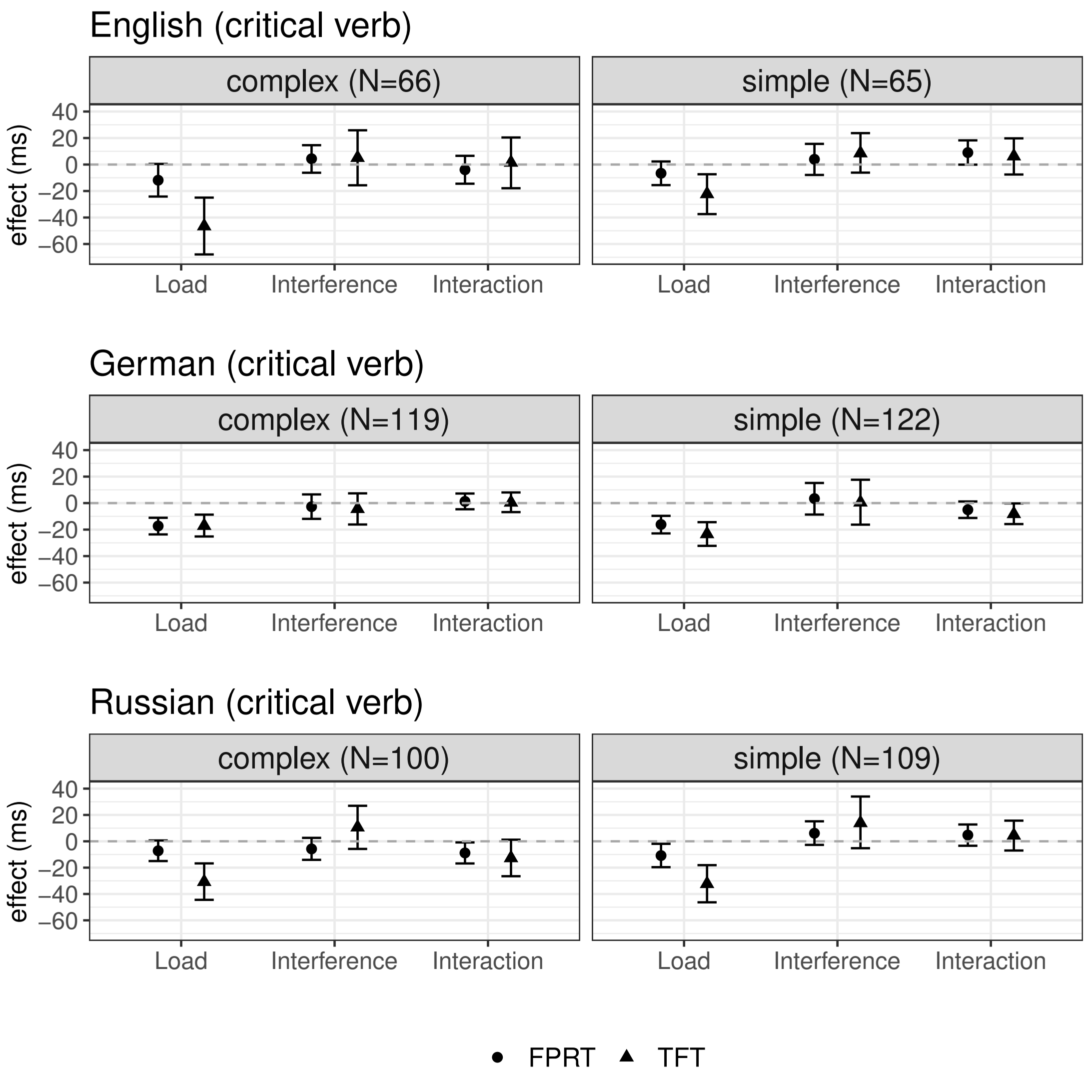


Figure 3.8: Effect of Load, Interference, and their interaction at the critical relative clause verb for the complex and the simple versions of the English, German, and Russian experiments. Values were back-transformed from the log scale to the millisecond scale. FPRT = first-pass reading times, TFT = total fixation times.

Before we turn to the Bayes factor results, note that we preregistered the Load × Interference prediction for TFT, predefining a 'region of practical equivalence' (or 'null region') to interpret the TFT results (e.g., Freedman et al., 1984; Hobbs & Carlin, 2008; Kruschke, 2015; Spiegelhalter et al., 1994). Because this null region is

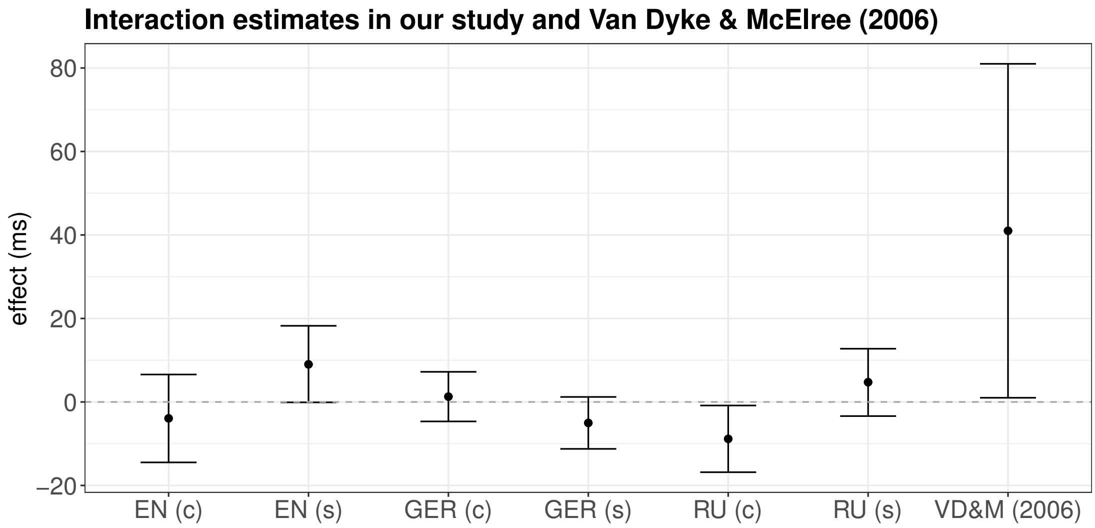


Figure 3.9: Interaction estimates with 95% credible intervals from our study (first-pass reading times), compared with interaction estimate with 95% confidence interval from Van Dyke & McElree (2006) (self-paced reading times). EN = English, GER = German, RU = Russian, (c) = complex version, (s) = simple version.

not suitable for FPRT, for clarity and brevity, we use Bayes factors to interpret the results of both reading measures here. The preregistration and the preregistered analysis alongside the code are available in the supplementary materials.

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 3.10 shows the Bayes factor results for Model 0, which does not include the interaction term, over Model 1, which does include the interaction term (this is denoted as BF_{01}). We show these Bayes factor results 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 for increasingly uninformative prior distribution standard deviations (SD) on the interaction (x-axis).

Most notably, the Bayes factor for English (simple, FPRT) shows some very modest evidence for the interaction in the predicted direction, at least under the more informative $\mathcal{N}(0, 0.05)$ prior. For German (simple, TFT) and Russian (complex, FPRT, TFT), the Bayes factors very weakly support an unexpected interaction under the more informative prior. For all other experiment versions, the Bayes factors suggest that the data are more likely to have been generated by Model 0, showing anecdotal to moderate evidence against a Load × Interference interaction.

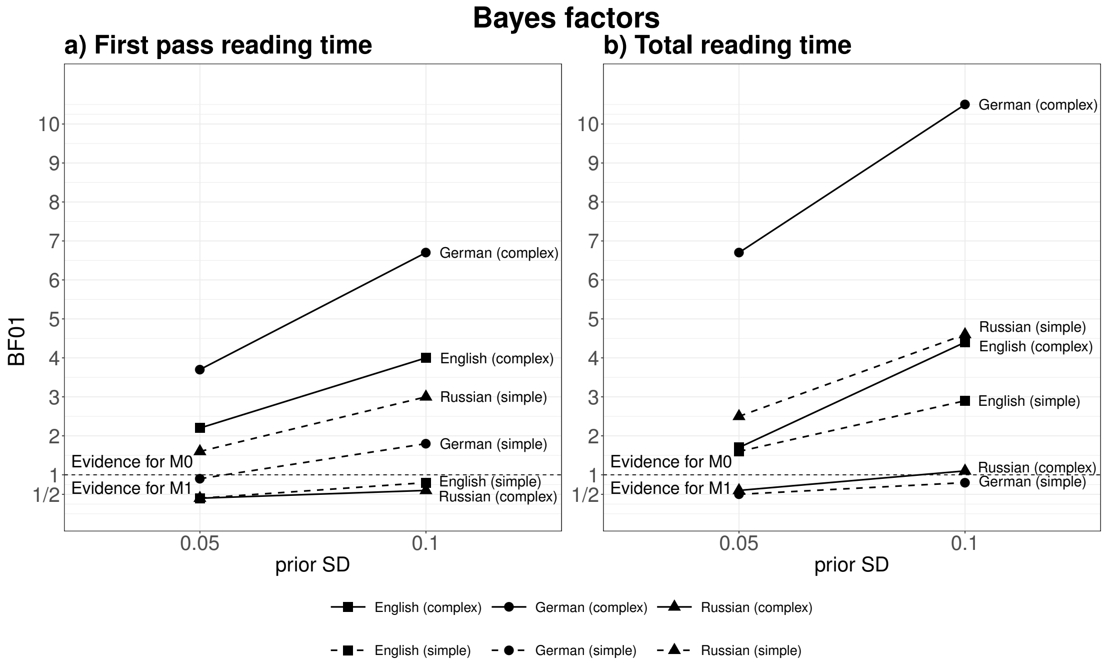


Figure 3.10: Bayes factor results for Model 0 over Model 1 (BF01) in a) first-pass reading time for the complex and the simple versions of the English, German, and Russian experiments. Bayes factor values are shown for increasingly uninformative prior distributions: $\mathcal{N}(0, 0.05)$, and $\mathcal{N}(0, 0.1)$

Effect estimates of the Language × Load × Interference interaction We further tested whether the Load × Interference interaction is modulated by language. The three-way Language × Load × Interference interaction indicates that there may be a difference in processing in English compared to German and Russian (Figure 3.11).

The influence of Depth of Processing The processing depth manipulation did not show the expected pattern in any language, namely, a smaller, or no interference effect in the simple compared with the complex experiment versions. In English, the reading time patterns were compatible with the idea that processing differs across the two version. However, unexpectedly, interference was observed only in the simple version. We therefore do not discuss the Processing

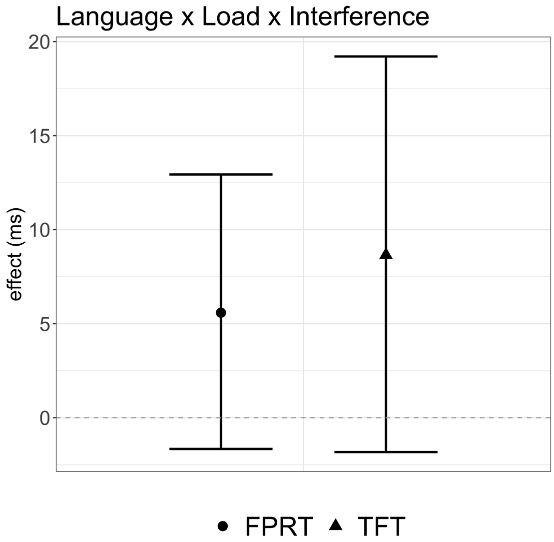


Figure 3.11: Shown are the posterior means and 95% credible intervals for the Language × Load × Interference interaction (first-pass reading times, total reading times). For the analysis, the combined data from all experiment versions had 304 subjects and 240 items. We tested whether the Load × Interference interaction in English differs from German and Russian. The contrasts for the factor Language were specified using the R package `hypr` (Rabe et al., 2020), testing the null hypothesis H0: EN $-$ (GER + RU)/2 = 0, i.e., whether the difference between the Load × Interference interaction in English (EN) vs. the interaction in German (GER) and Russian (RU) (averaged) is equal to zero. English was coded as $+0.66$, and German and Russian as -0.33.

depth × Load × Interference interactions further; these can be inspected in the supplementary materials.

3.2.5 Discussion

This study re-investigated proactive cue-based retrieval interference in sentence comprehension. Our large-sample cross-linguistic study only shows anecdotal evidence for the predicted interaction in one language, version and reading measure. At best, we can conclude that the Bayes factors very weakly support the expected interaction in English (simple version). For all other experiment versions, the Bayes factors revealed moderate evidence against the predicted interaction.

Overall, our data do not show much support for cue-based retrieval interference from sentence-external distractors. Before drawing conclusions, it is important to rule out task-related explanations for these results. 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 low 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 clear 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 D). For the simple version of the English experiment, the interaction estimate has a slightly wider credible interval ranging from -5 to 17 ms. These results indicate that even when all memory nouns are encoded in memory, they do not seem to interfere with within-sentence dependency resolution in any of the other experiment versions.

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 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 D). 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 D.2). This 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.

Furthermore, 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. It was reasoned that items in focus have more distinctive memory representations which may reduce the magnitude of the interference effect. Conversely, it could be argued 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.

3.3 General Discussion

The main goal of this study was to re-investigate cue-based theories' prediction of cue-dependent retrieval interference. The present study is the first to test proactive similarity-based retrieval interference cross-linguistically, using the same method, design, and comparable sentence structures across languages. We aimed to extend previous findings on English object clefts to the online processing of filler-gap dependencies, using the dual-task paradigm by Van Dyke and McElree (2006). In addition, the relatively large sample sizes in German and Russian compared to most previous interference studies provide more precise estimates of the interference effect. A further novel contribution of this study is a within-subjects manipulation of processing depth, investigating proactive interference under varying task demands.

Our analyses only showed some weak evidence in favor of the predicted Load $\times$ Interference interaction in the simple version of the English experiment. Thus, only the English data lend any support to the hypothesis that semantically similar, sentence-external distractors in memory can interfere with retrieval during real-time sentence comprehension. In contrast to English, 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.

The weak evidence for interference in English and evidence against interference in German and Russian may point to cross-linguistic processing differences. In German and Russian the accusative case marking of the grammatical object may increase distinguishability of the retrieval target, and therefore eliminate interference during online dependency formation. Whether these differences are due to case marking attenuating interference is speculative, and requires further investigation, using a cross-linguistic design that explicitly manipulates case marking.

If proactive interference occurs in English but not in German or Russian, then it is surprising that interference was only observed in English when simple questions were asked but not when complex questions were asked. Our depth of processing hypothesis postulated that inducing shallow processing would lead to less or no interference during online sentence comprehension (Logačev & Vasishth, 2016; Swets et al., 2008). However, only the simple version showed some support for proactive interference. In the complex version, the effect was not replicated. Here, we refer to replication as observing overlapping effect estimates with the predicted positive sign. What could be the reason for not replicating the effect in the complex conditions, despite testing the same participants across the two experiment versions? One explanation might be that this effect only arises under good-enough processing. Laurinavichyute and von der Malsburg (2021a) recently found support for agreement attraction effects under superficial but not deep processing conditions. Our study might also reveal this difference in processing depth. Furthermore, in our study, it is plausible that the difference in processing depth does not reflect superficial vs. deep processing as 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 to this proposal. 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. The interaction estimate in the simple version of Russian (vs. the complex version) may also point towards this heightened attention hypothesis, although the Bayes factors in this study offered no evidence for the interference effect in the simple version of Russian.

An alternative explanation for observing an interference effect in English under low but not high task demands is that this result is due to noise in the data. Given that—of all experiment versions—only the lower-sample size English study (simple version, FPRT) showed weak evidence for proactive interference, this could be an accidental outcome that may not be replicable in a future study. To show that proactive retrieval interference is replicable under a low processing load, a confirmatory experiment would have to be conducted (e.g., de Groot, 1956/2014; Nicenboim et al., 2018; Vasishth et al., 2018).

If cue-based parsing mechanisms are an integral part of language processing, it is interesting that our study does not show cross-linguistic support for it. Why don't we observe the predicted reading time pattern, with one exception? While cross-linguistic processing differences and processing depth might play a role, other factors could have contributed to our results. 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 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 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 likely to find ourselves in a relatively low power scenario, as Figure 3.1 demonstrated. Even with larger participant samples, as in German or Russian, such small effects may remain undetected. Consequently, to detect 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. For example, Van Dyke and McElree (2011) showed in 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 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, particularly when the distractor intervenes linearly between the retrieval target and retrieval site. For example, the retrieval cues used to reactivate the filler noun phrase may strongly distinguish within-sentence material from other material in working memory. This theoretical possibility has been suggested by authors who proposed that 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.

The findings from our cross-linguistic large-sample study allow for limited conclusions. In the psychological sciences, an argument has been made for more cautious conclusions, as single studies frequently overstate the generality of effects (Simons et al., 2017; Yarkoni, 2020). We also favor this approach in the reporting of psycholinguistic research results.

In sum, our study provides only weak evidence that proactive retrieval interference from sentence-external material can interfere with within-sentence dependency resolution in English filler-gap dependencies. This finding is compatible with cue-based retrieval accounts, and with the reading time pattern observed in (Van Dyke & McElree, 2006). Our data contrast the previous research because in our study, the effect was only observed under a low task demand. Two further languages, German and Russian, revealed either equivocal evidence, or evidence against the predicted interference effect, possibly hinting at between-language processing differences that cue-based theories do not currently account for. This is subject to further investigation. Until further data become available, our findings suggest that proactive retrieval interference from sentence-external material arises only in very specific linguistic contexts.

3.4 Conclusions

Our data indicate that extra-sentential items encoded in memory can, but may not always, interfere with within-sentence dependency formation. The only weak evidence comes from English sentence comprehension under low task demands. This finding—at least weakly—suggests that the proactive interference of the type shown here has its source at retrieval, indicating that a cue-based mechanism—using semantic cues for retrieval—is in play during real-time dependency formation. Our study did not find cross-linguistic support for proactive interference, suggesting that language-specific features such as distinctive case marking may reduce interference effects.

Appendix B

Examples of German and Russian experimental materials

B.1 German stimuli

We created 40 experimental items for the complex and 40 new items for the simple version of the experiment. Table B.1 shows two example items; B.1A is followed by a complex comprehension question, and B.1B is followed by a simple comprehension question. For B.1A, 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 3.3, 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 3.3). 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 B.1. 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.

Table B.1: German example items (complex and simple version).

A) Memory load:					Kühlschrank			Waschmaschine			Computer		
					fridge			*washing machine*			*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 <u>*trank*</u>, 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 <u>*roch*</u>, 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 <u>*trank*</u>, 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 <u>*roch*</u>, 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?)*

B.2 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 B.2 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).

Table B.2: Russian example items (complex and simple version).

A) Memory load:				ковер		подушка		покрывало	
				carpet		pillow		cover	
No interference									
То	кресло,	которое	старушка,	любящая	антиквариат	<u>сломала</u>	недавно,	относится к	…
The	armchair,	that	elderly woman,	loving	antiques	broke	recently,	belongs to	…
Interference									
То	кресло,	которое	старушка,	любящая	антиквариат	<u>присмотрела</u>	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	<u>сломала</u>	недавно,	относится к	…
The	armchair,	that	elderly woman,	loving	antiques	broke	recently,	belongs to	…
Interference									
То	кресло,	которое	старушка,	любящая	антиквариат	<u>присмотрела</u>	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, <u>прокусила</u> 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, <u>заметила</u> 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, <u>прокусила</u> 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, <u>заметила</u> 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?

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.

Appendix C

Description of filler items

C.1 English

The English study had 90 filler items. We created 50 "true" filler items of two varying syntactic structures 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.

C.2 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.

C.3 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 D

Perfect recall data

D.1 Comprehension question accuracy in perfect recall trials

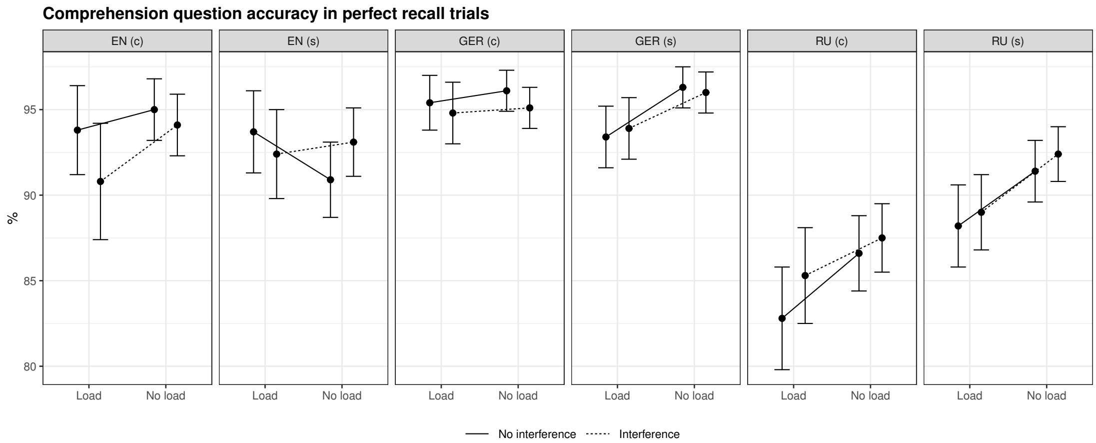


Figure D.1: By-condition mean comprehension question accuracies (in percent) with 95% confidence intervals in perfect recall trials. EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version

D.2 Raw by-condition reading times in perfect recall data

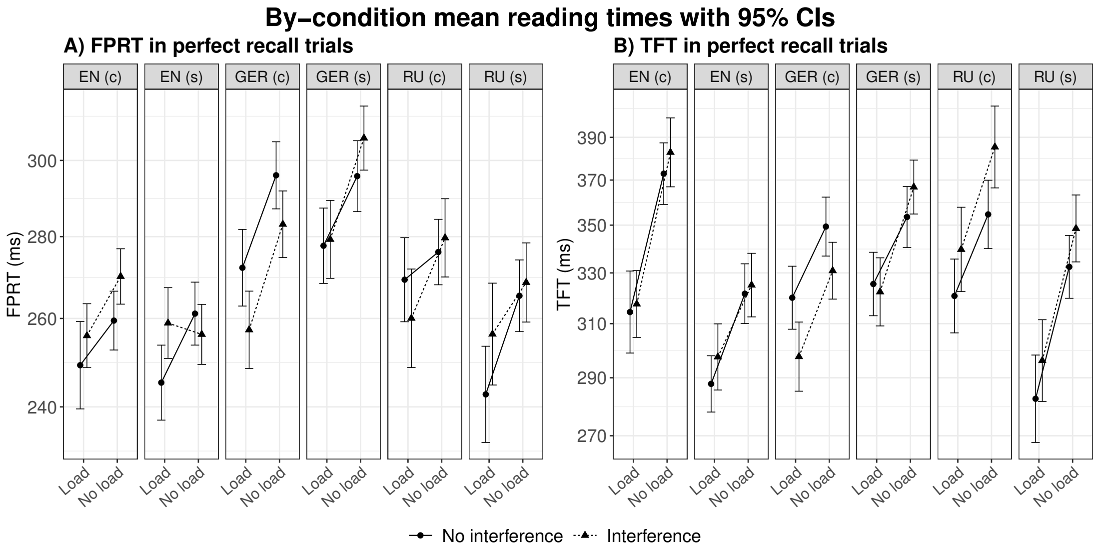


Figure D.2: By-condition mean reading times with 95% confidence intervals. In (A) and (B), FPRT and TFT are shown for trials with perfect recall, i.e., recall of three memory nouns in the correct order. EN = English, GER = German, RU = Russian; (c) = complex version, (s) = simple version.

D.3 Reading time analyses for data with perfect recall

Study	Load	Interference	Interaction
	FPRT Posterior mean [95% credible intervals]		
English, complex	−12 ms [−27, 3]	4 ms [−8, 17]	−4 ms [−16, 8]
English, simple	−7 ms [−18, 3]	1 ms [−11, 13]	6 ms [−5, 17]
German, complex	−17 ms [−24, −10]	−4 ms [−13, 5]	1 ms [−6, 8]
German, simple	−16 ms [−24, −8]	6 ms [−7, 20]	−3 ms [−10, 4]
Russian, complex	−8 ms [−18, 1]	−4 ms [−13, 5]	−7 ms [−15, 2]
Russian, simple	−14 ms [−24, −5]	7 ms [−3, 16]	5 ms [−4, 15]

Table D.1: 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.

Study	Load	Interference	Interaction
	TFT Posterior mean [95% credible intervals]		
English, complex	−51 ms [−74, −28]	6 ms [−19, 32]	3 ms [−21, 27]
English, simple	−26 ms [−44, −9]	1 ms [−15, 18]	−1 ms [−17, 15]
German, complex	−19 ms [−29, −9]	−3 ms [−15, 8]	1 ms [−8, 10]
German, simple	−26 ms [−38, −15]	4 ms [−15, 23]	−7 ms [−16, 2]
Russian, complex	−32 ms [−47, −18]	13 ms [−3, 30]	−11 ms [−27, 5]
Russian, simple	−42 ms [−56, −28]	9 ms [−11, 29]	0 ms [−12, 13]

Table D.2: 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.

Chapter 4

Synthesis of syntactic and semantic similarity-based interference effects

In chapters 2 and 3, we presented the findings of two larger-sample studies on syntactic and semantic interference effects in online sentence comprehension. The experiments in chapter 2 examined the time course of retroactive syntactic and semantic interference effects cross-linguistically. This study suggested that both types of interference can occur in parallel.

The interference study in chapter 3 investigated proactive semantic, cue-based retrieval interference from sentence-external distractors cross-linguistically. This study showed only very modest evidence for semantic interference in English under superficial processing. However, German and Russian showed evidence against this effect.

Next, we ask what we can learn from the combined results of our own experiments and the original Van Dyke experiments. We therefore synthesize the results from the experiments in chapters 2 (retroactive interference) and 3 (proactive interference) and the experiments by Van Dyke and colleagues by computing meta-analysis estimates.

We first present the meta-analysis estimates for syntactic and semantic interference effects from within-sentence distractors, as discussed in chapter 2. We included the data from our English and German experiments, and the estimates from the original reading experiments in Van Dyke (2007), as well as Van Dyke and Lewis (2003), and Van Dyke and McElree (2011). The Van Dyke (2007) sentences were shown in example (4) of chapter 2. The Van Dyke and Lewis (2003) study used the two conditions of these sentences that included the syntactic manipulation. Van Dyke and McElree (2011) used comparable sentence structures (see example 11).[1]

(11) a. The judge who had declared that *the motion* was inappropriate realized that **the attorney** in the case **compromised** during the negotiations.
b. The judge who had declared that *the witness* was inappropriate realized that **the attorney** in the case **compromised** during the negotiations.
c. **The attorney** who the judge realized had declared that *the motion* was inappropriate **compromised** during the negotiations.
d. **The attorney** who the judge realized had declared that *the witness* was inappropriate **compromised** during the negotiations.

For our own experiments, we extracted the effect estimates from the Bayesian linear mixed-effects models. For the studies by Van Dyke and colleagues, we extracted the effect estimates of each comparison from the reported raw reading times, and used the largest SE reported alongside the condition means, following Jäger et al. (2017).

One challenge for eye-tracking data is that multiple eye-tracking measures could be used for the meta-analysis. For example, in our experiments of chapter 2 (retroactive interference), reading time slowdowns were observed in first-pass reading times, regression-path durations, and total fixation times. This was also the case in the Van Dyke (2007) experiments. Indeed, many previous eye-tracking studies on interference report effects in a variety of measures. A non-exhaustive overview is given in Table 4.1.

[1] Note that conditions (a,b) in Van Dyke and McElree (2011) tested proactive interference. We added these here because they also involved within-sentence distractors, although this type of manipulation might be weaker, as discussed earlier.

Table 4.1: Similarity-based interference effects by eye-movement: Shown are the eye-tracking measures that were reported to show significant effects in experiments from 15 published interference studies. We show the proportion of significant effects for each reading measure, alongside the published experiments. The overall count of effects that were reported to be significant was 44. Of these, 11 of the 44 significant effects were reported in TFT (i.e., .25). The following reading measures are listed: (reg-cont) FFD = (regression-contingent) first fixation duration, FPRT = first-pass reading time, FPR = proportion of first-pass regressions out, RBRT = right-bounded reading time, Reg in = Regressions in, Cumu RT = cumulative reading times, RPD = regression-path duration, RRT = re-reading time, SPT = second-pass time, TFT = total fixation time.

Publication	**ET measure**	**Proportion**
Acuña–Fariña et al. (2014)	(reg-cont) FFD	.07
Jäger et al. (2015)		
Patil et al. (2016)		
Acuña–Fariña et al. (2014)	cumu RT	.02
Cunnings and Felser (2013) E2	**FPRT**	.20
Cunnings and Sturt (2014) E1		
Jäger et al. (2015) E1, E2		
King et al. (2012)		
Pearlmutter et al. (1999) E2		
Van Dyke (2007) E2 LoSyn, E2 LoSem, E3 LoSem		
Acuña–Fariña et al. (2014)	FPR	.09
Patil et al. (2016)		
Pearlmutter et al. (1999) E2		
Van Dyke (2007) E2 LoSem		
Acuña–Fariña et al. (2014)	Reg in	.02
Jäger et al. (2015) E1, E2	RBRT	.05
Cunnings and Sturt (2018) E2	**RPD**	.20
Felser et al. (2009) E2b		
Jäger et al. (2015) E1, E2		
Parker and Phillips (2017) E3		
Van Dyke (2007) E2 LoSyn, LoSem, E3 LoSyn, LoSem		

Continuation of Table 4.1		
Acuña–Fariña et al. (2014) Cunnings and Sturt (2018) E1, E2 Dillon et al. (2013) E1 agr Jäger et al. (2015) E2 Parker and Phillips (2014) E1, E2 Pearlmutter et al. (1999) E2 Van Dyke (2007) E2 LoSem Van Dyke and McElree (2011) E1b pro, retro	**TFT**	.25
Sturt (2003) E1 Parker and Phillips (2017) E1, E2, E3	RRT/SPT	.09

Because effects have been reported in multiple measures, researchers may disagree on which measures should be used to compute a meta-analysis estimate of similarity-based interference. To account for this, we synthesize the data for the three most commonly reported eye-tracking measures, first-pass reading times, regression-path durations, and total fixation times. These three measures also showed effects in our own studies, as well as Van Dyke (2007). We also include the estimates from self-paced reading experiments (Van Dyke and Lewis, 2003, Experiment 4; Van Dyke, 2007, Experiment 1).

Figure 4.1 shows the full posterior distributions for the syntactic and semantic interference effects. The top panel shows the posteriors for the *syntactic* interference effect in A) first-pass reading times and self-paced reading times, B) regression-path durations and self-paced reading times, and C) for total fixation times and self-paced reading times. The bottom panel shows the posteriors for the *semantic* interference effect.

Figure 4.2 shows the posterior means with their 95% credible intervals of the meta-analysis estimates, next to the effect estimates from the individual studies that entered the meta-analysis. For the meta-analysis estimate of syntactic interference, these were twelve pairwise comparisons from reading studies by Van

Dyke (2007) and Van Dyke and Lewis (2003), as well as from our experiments presented in chapter 2 (retroactive interference). The top panels visualize the *syntactic* interference effects for first-pass reading times and self-paced reading times, regression-path durations and self-paced reading times, and for total fixation times and self-paced reading times. The bottom panels of Figure 4.2 show the *semantic* interference effects. Again, we show the meta-analysis estimates next to the estimates from each individual comparison that entered the analysis. For the meta-analysis estimate of semantic interference, fourteen comparisons from reading studies by Van Dyke (2007) and Van Dyke and McElree (2011), and our experiments presented in chapter 2 (retroactive interference) entered the analysis.

For syntactic interference, the meta-analysis estimate has a positive sign across the three reading measures. In first-pass reading times, the meta-analysis estimate indicates that the most plausible values of the effect range from -2 ms to 19 ms, in regression-path durations, they range from 12 ms to 60 ms, and in total fixation times, they range from 8 ms to 42 ms. This suggests that a reading time slowdown can occur at the point of retrieval when syntactically similar distractors interfere within the sentence.

For the semantic interference effect, the first-pass reading time meta-analysis estimate is centered around zero (95% CrI $[-5, 6]$ ms). The regression-path duration and total fixation time estimates both have a positive sign. In regression-path duration, it ranges from -1 ms to 18 ms, and in total fixation times, it ranges from 2 ms to 29 ms. This suggests that semantically similar distractors can cause a processing slowdown at the point of retrieval, although there is no indication that semantic interference occurs during the first pass, that is, in the earliest moments of processing.

We now turn to the meta-analysis of *proactive* semantic cue-based retrieval interference, as discussed in chapter 3. For this, we included the data from our dual-task experiments (the simple and complex versions of English, German, Russian), as well as the Van Dyke and McElree (2006) and Van Dyke et al. (2014) estimates. Recall that the effect of interest for the dual-task studies is the interaction be-

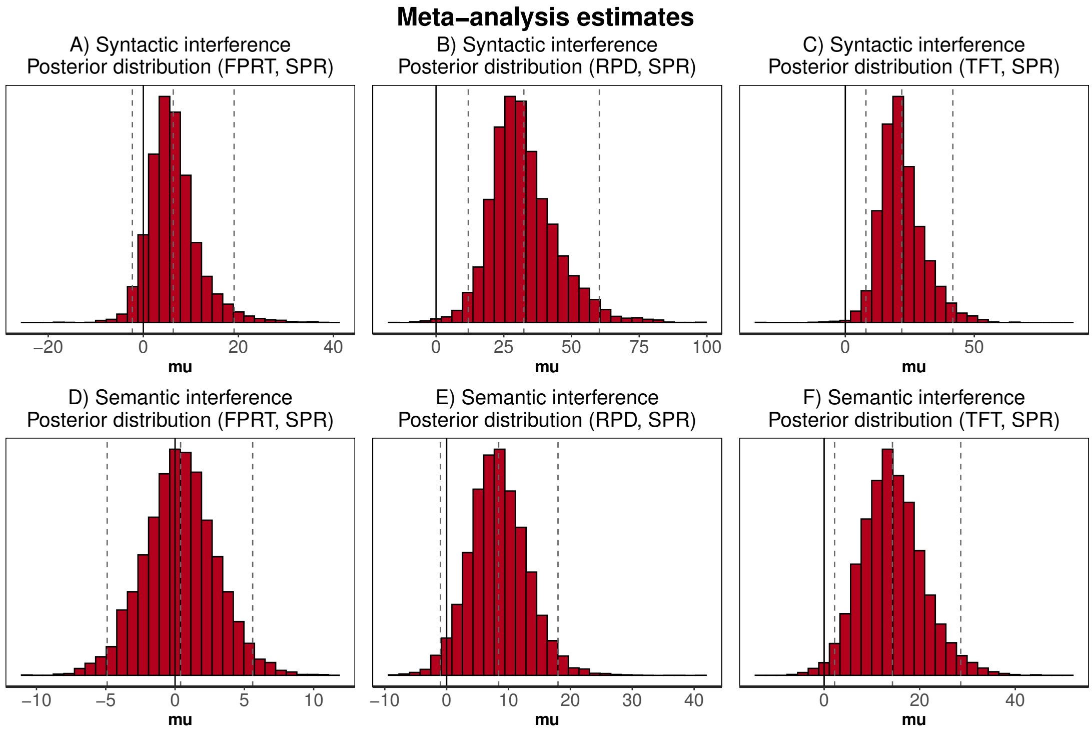


Figure 4.1: The top panel shows posterior distributions of the syntactic interference effect for A) FPRT and SPR, B) RPD and SPR, C) TFT and SPR.
The bottom panel shows the posterior distributions for the semantic interference effect. Dashed lines show the posterior means and 95% credible intervals.

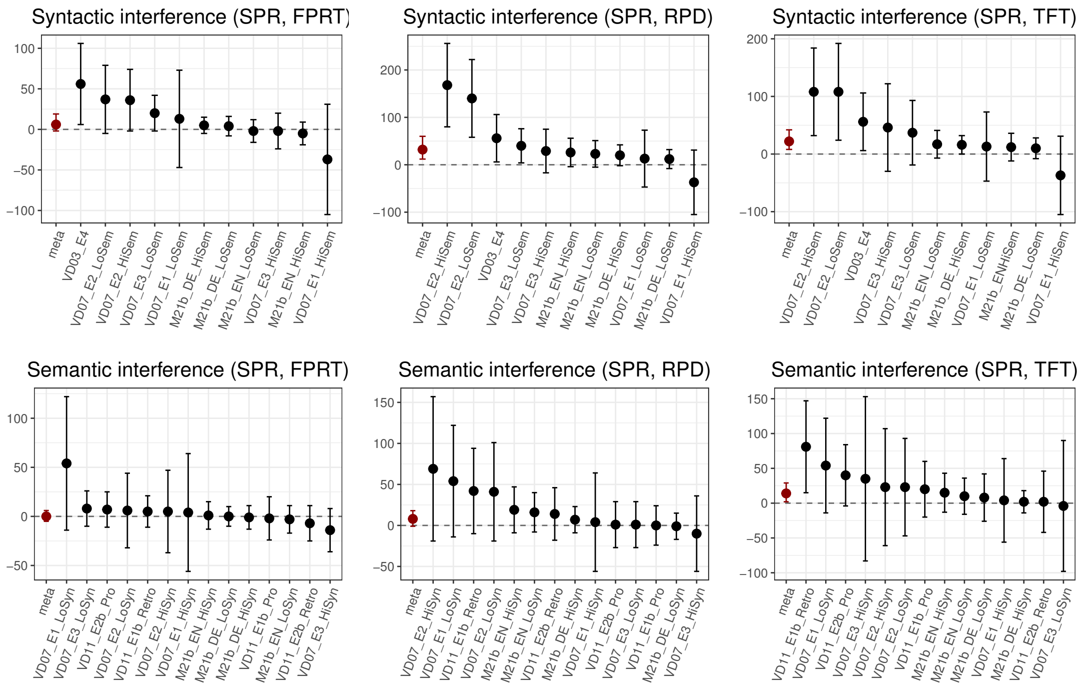


Figure 4.2: Shown are the meta-analysis estimates (posterior means and 95% credible intervals) next to the individual effect estimates (means and 95% confidence intervals) of the pairwise comparisons that entered the meta-analysis. meta = meta analysis estimate, VD refers to the studies by Van Dyke et al., M refers to the experiments in this dissertation (Mertzen et al.), LoSyn/HiSyn = Low and High syntactic interference conditions, LoSem/HiSem = Low and High semantic interference conditions. EN = English, DE = German

tween Load and Interference. Here, we do not include the estimates from Gordon et al. (2002) and Fedorenko et al. (2006) because these studies do not explicitly test cue-based retrieval interference (their results can be explained by encoding interference).

Recall that for the experiments in chapter 3 (proactive interference), only the simple version of the English experiment showed some weak evidence of the expected effect in first-pass reading times. For our experiments, we thus extracted the effect estimates for first-pass reading times from the Bayesian linear mixed-effects models. For the Van Dyke and McElree (2006) study, we extracted the Load $\times$ Interference interaction estimate from the reported raw reading times. As the standard error (SE), we used the largest SE reported alongside the raw condition means (Jäger et al., 2017). For Van Dyke et al. (2014), the Load $\times$ Interference interaction estimate and the SE were taken from their Table 4 (p. 46).

Figure 4.3 shows the posterior distribution for the proactive semantic interference effect from sentence-external distractors. Figure 4.4 shows the 95% credible intervals of the meta-analysis estimates alongside the estimates from the individual experiments that entered the meta-analysis.

The meta-analysis estimate indicates that the most plausible values of the interaction range from -7 ms to 8 ms. Overall, we see no indication of a Load $\times$ Interference interaction, that is, the meta-analysis offers no support for the hypothesis that semantically-similar, sentence-external distractors can interfere with within-sentence dependency resolution.

In sum, the meta-analysis estimates for proactive semantic interference of the kind discussed in chapter 3 show no indication that sentence-external distractors interfere with within-sentence dependency resolution. On the other hand, the meta-analysis estimates for syntactic and semantic interference from within-sentence distractors of the kind discussed in chapter 2 show some indication that syntactically and semantically similar distractors cause a reading time slowdown during retrieval.

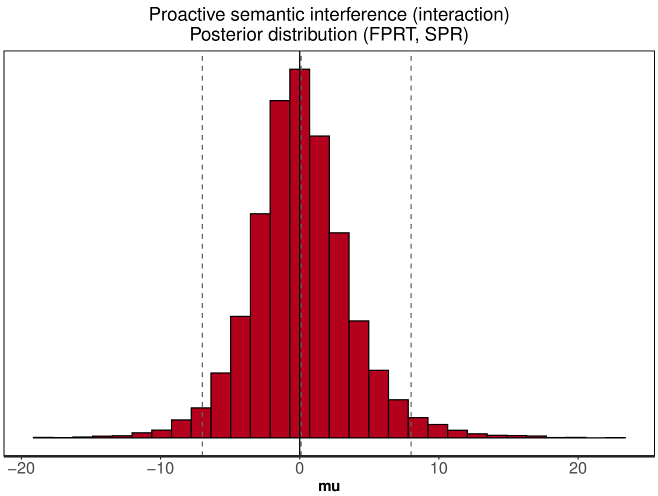


Figure 4.3: Posterior distribution of the proactive semantic interference effect for FPRT and SPR. The interaction estimates from Van Dyke & McElree (2006) and Van Dyke et al. (2014), as well as the interaction estimates from our experiments in Chapter 3 (English complex, simple; German complex, simple; Russian complex, simple) entered the meta-analysis. Dashed lines show the posterior mean and 95% credible intervals.

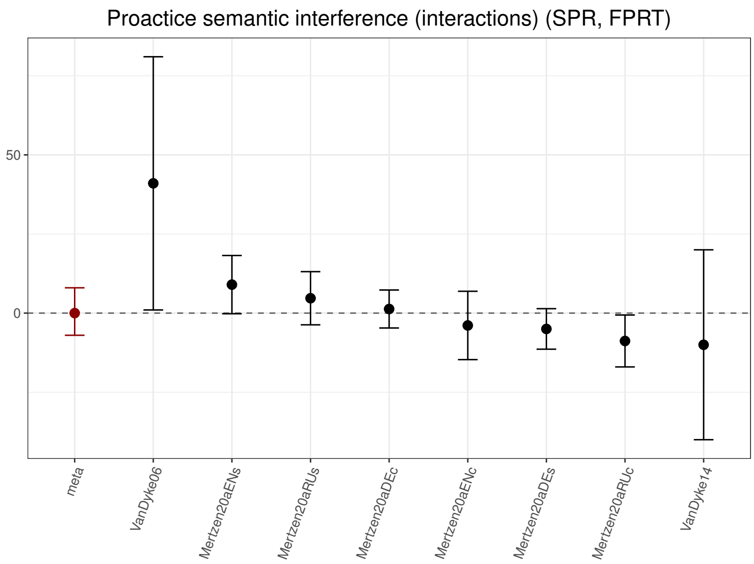


Figure 4.4: Meta-analysis estimate (posterior mean and 95% credible interval) next to the interaction estimates (means and 95% confidence intervals for the Van Dyke et al. studies; posterior means and 95% credible intervals for the Mertzen et al. studies) that entered the meta-analysis. Meta = meta-analysis estimate, VanDyke06 = Van Dyke & McElree (2006), VanDyke14 = Van Dyke, Johns, & Kukona (2014), Mertzen20a refers to the studies in chapter 3, EN = English. DE = German, RU = Russian, c = complex version, s = simple version.

Our results from two larger-scale cross-linguistic studies (chapters 2, 3), and the meta-analysis results in this chapter, might suggest that interference is modulated by the position of the distractor. Our empirical data and the meta-analysis estimates suggest that when distractors occur within the sentence, specifically, when they intervene between the verb and target noun phrase,[2] they can cause similarity-based interference.

By contrast, when a distractor precedes a target dependency, specifically when it occurs sentence-externally, it might not interfere, or at least interfere less. It has previously been reported that proactive interference, that is, interference from distractors that precede a target dependency, might be weaker than retroactive interference effects—even when the distractors occur within the sentence (Jäger et al., 2017; Van Dyke & McElree, 2011). Proactive interference effects from sentence-external distractors may therefore be too subtle to be detected in reading studies, even with a relatively large participant sample.

[2] Although not exclusively, as there are a number of studies that have found proactive interference effects where the distractors occurred within the critical sentence (e.g., Jäger et al., 2015; Parker and Phillips, 2014; Sturt, 2003; Van Dyke and McElree, 2011).

Chapter 5

Conclusion

In this dissertation, I presented a cross-linguistic investigation of similarity-based interference effects in sentence comprehension. The studies probed the constraints involved in the online resolution of long-distance dependencies. Specifically, we asked whether we can see support for syntactic and semantic information being used during the formation of long-distance dependencies. We also asked whether both types of information can be used simultaneously.

Chapter 2 investigated the time course of syntactic and semantic interference during online dependency formation. Two eye-tracking-while-reading studies tested English and German subject-verb dependencies, using the Van Dyke (2007) design. The findings in our experiments partially align with the findings in Van Dyke (2007): The cross-linguistic data from our studies on retroactive interference supports the hypothesis that both syntactically and semantically-similar items can cause similarity-based interference. In addition, our study suggests that the parser can use both types of retrieval cues simultaneously. Such simultaneous effects were not observed in Van Dyke (2007).

Chapter 3 re-examined the Van Dyke and McElree (2006) claim of semantic cue-based retrieval interference from sentence-external distractors on within-sentence dependency resolution. It extended the previous work in several ways: First, we conducted a cross-linguistic test of cue-based retrieval interference. We tested English, German and Russian filler-gap dependencies with the Van Dyke and McElree (2006) dual-task design, using the eye-tracking-while-reading method. Particularly the German and the Russian studies had larger sample sizes than previous studies on interference, yielding much more precise effect estimates than previous work. In addition, interference was tested for different processing depths. Our results are partially in line with results by Van Dyke and McElree (2006). The empirical data presented in chapter 3 suggests that there is—at best—some very weak evidence for proactive interference from sentence-external distractors in English, but only under a low task demand. In both German and Russian, we found evidence against proactive interference from sentence-external distractors. The empirical data from our studies on proactive interference only allow for limited conclusions: Given the extremely weak support for an effect in only one language, experiment version, and reading measure, this result could be accidental, and may indicate that the parser does not consider semantically similar, sentence-external items as potential retrieval targets. Alternatively, it is possible that cue-based retrieval from sentence-external distractors arises only in specific linguistic contexts. In order to validate the exploratory finding from English, a highly powered, confirmatory replication study is needed.

In Chapter 4, the data from our experiments and the Van Dyke studies were synthesized. The meta-analysis largely supports the conclusions presented in chapters 2 and 3. The meta-analysis estimates for similarity-based interference effects from sentence-internal distractors showed reading time slowdowns when syntactically or semantically similar distractors interfered, supporting the results presented in chapter 2. By contrast, the meta-analysis estimate for proactive interference showed no indication of a reading time slowdown for semantically similar, sentence-external distractors. Thus, the meta-analysis shows no support for the hypothesis that sentence-external distractors interfere with sentence-internal dependency resolution.

Overall, what have we learned from the experiments presented in this dissertation? Our larger-scale, cross-linguistic work shows support for both syntactic and semantic similarity-based interference effects in online sentence comprehension—at least from within-sentence distractors. The simultaneous effects of syntactic and semantic interference support the hypothesis that, in principle, structural and non-structural cues can be used simultaneously. Given the post-critical semantic effects (in German), and the lower comprehension question accuracy when the target and distractor noun were animate, it is possible that semantic interference may have a more lasting, detrimental effect on overall comprehension.

Taken together, the experiments in this dissertation might indicate that interference effects are contingent on the position of the distractor. Our data in chapter 2 (retroactive interference) show that interference arises from within-sentence distractors. On the other hand, the data in chapter 3 (proactive interference) do not show much support for interference from sentence-external distractors. Our synthesis of the data in chapter 4 (that includes all our data from three languages, as well as the Van Dyke and McElree (2006) and Van Dyke et al. (2014) estimates) also does not show any indication of the predicted effect. One possibility is that distractors may predominantly cause processing difficulty when they appear sentence-internally. We discussed in chapter 3 (proactive interference) the proposal that retrieval is guided by clause-bounding retrieval cues (Wagers et al., 2009). There is also some data that is compatible with such a proposal, namely, the finding that retrieval interference is reduced for distractors inside appositive relative clauses which could be viewed as parentheticals that are distinct from the target sentence (Dillon et al., 2017). If the effect we observed in English is spurious, it is also possible that the parser does not consider material that is not integrated in the syntactic structure during retrieval.

In addition to the possibility that distractors predominantly interfere when they occur sentence-internally, it is possible that distractors cause greater processing difficulty when they intervene between the target and retrieval point (i.e., in a retroactive configuration), as opposed to prior to the critical dependency (i.e., in a proactive configuration). There is some support for this hypothesis in Van Dyke and McElree (2011). If distractors cause greater interference when they occur sentence-internally, and retroactively, then proactive interference from sentence-external materials likely are very subtle and hard to detect in our standard reading studies.

For the proactive investigation in chapter 3, if the result in the simple version of the English experiment was not an accidental outcome, then this may show that proactive interference effects from sentence-external distractors only arise in specific linguistic contexts. Firstly, we did not see cross-linguistic evidence for the effect; we only observed an effect in English. If sentence external distractors show an effect in English, but not in German and Russian, then what may cause this cross-linguistic difference? Three larger-sample plausibility rating experiments (English, German, Russian) checked that divergent plausibility ratings are not a potential confound in the experimental materials. It is possible that the overt case marking in German and Russian makes items in memory more distinguishable, diminishing or eliminating proactive interference effects, and hence, we only observe the effect in English. This is subject to further investigation.

Second, we only observed proactive interference under a low task demand. We discussed in chapter 3 (proactive interference) that the superficial processing manipulation may not, in fact, have induced superficial processing, but simply "typical" (attentive) processing. Even though the simple questions allowed participants to underspecify syntactic relations, they may not have done so. If so, then proactive interference may arise in typical reading, but it might disappear when the comprehension questions require the reader to direct their focus toward establishing, and memorizing, all dependencies in order to give a correct response.

Finally, while the studies in this dissertation show support for similarity-based interference in sentence comprehension, for our experiments in chapter 2 (retroactive interference) there is some uncertainty with regard to the source of the interference effects, that is, whether they are indeed cue-based retrieval interference effects. Particularly, because interference effects were observed pre-critically, we can only cautiously conclude that the effects are the result of cue-based retrieval interference. Only a highly-powered study with an improved design can shed more light on this issue. The study in chapter 3 (proactive interference) showed only modest evidence for a cue-based retrieval interference effect in one language, in one eye-tracking measure, under a low task demand. This result could be due to noise in the data. Thus, as for the study in chapter 2, an additional highly-powered study would be required to further consolidate the evidence for this effect in English sentence comprehension. We are currently planning to conduct a highly-powered replication study of Van Dyke and McElree (2006), and our experiments reported in chapter 3 on proactive interference.

References

Acuña–Fariña, J. C., Meseguer, E., & Carreiras, M. (2014). Gender and number agreement in comprehension in Spanish. *Lingua*, *143*, 108–128. https://doi.org/10.1016/j.lingua.2014.01.013 (cit. on pp. 103, 104)

Aoshima, S., Phillips, C., & Weinberg, A. (2004). Processing filler-gap dependencies in a head-final language. *Journal of Memory and Language*, *51*(1), 23–54. https://doi.org/10.1016/j.jml.2004.03.001 (cit. on p. 40)

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 (cit. on pp. ii, 2, 10, 14, 34, 38, 48)

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 (cit. on pp. 54, 55)

Bader, M., & Lasser, I. (1994). Perspectives on sentence processing. In C. Clifton, L. Frazier, & K. Rayner (Eds.). Lawrence Erlbaum Associates. (Cit. on p. 40).

Birch, S., & Rayner, K. (1997). Linguistic focus affects eye movements during reading. *Memory & Cognition*, *25*, 653–660. https://doi.org/10.3758/bf03211306 (cit. on p. 58)

Brasoveanu, A., & Dotlačil, J. (2020). Syntax as a cognitive process: Left-corner parsing with visual and motor interfaces. *Computational cognitive modeling and linguistic theory* (pp. 57–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-31846-8_4. (Cit. on p. 39)

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28 (cit. on pp. 61, 73).

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 (cit. on p. 61)

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(1), 77–94. https://doi.org/10.1017/S0140525X99001788 (cit. on p. 49)

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01 (cit. on pp. 21, 70)

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 University Press. (Cit. on p. 58).

Chomsky, N. (1986). *Barriers*. MIT Press. (Cit. on p. 38).

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 (cit. on pp. 3, 12, 51)

Clark, H. H., & Begun, J. S. (1971). The semantics of sentence subjects. *Language and Speech*, *14*(1), 34–46. https://doi.org/10.1177/002383097101400105 (cit. on p. 39)

Clifton, C., Juhasz, B., Ashby, J., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language*, *49*, 317–334. https://doi.org/10.1016/B978-008044980-7/50017-3 (cit. on pp. 2, 8)

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. Van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–372). Elsevier. (Cit. on p. 21).

Cunnings, I., & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, *28*(1-2), 188–219. https://doi.org/10.1080/01690965.2010.548391 (cit. on pp. 54, 103)

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 (cit. on pp. iii, 2, 10, 36, 48, 54, 103)

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 (cit. on pp. ii, 2, 10, 34, 48, 54, 103, 104)

de Groot, A. (1956/2014). 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 Psycho-*

logica, *148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001 (cit. on p. 85)

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 (cit. on pp. 86, 113)

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 (cit. on pp. iii, 2, 10, 34, 36, 48, 54, 104)

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *43*(12), e12800. https://doi.org/10.1111/cogs.12800 (cit. on pp. 9, 10)

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 (cit. on pp. 2, 48, 49, 108)

Felser, C., Sato, M., & Bertenshaw, N. (2009). The on-line application of Binding Principle A in English as a second language. *Bilingualism: Language and Cognition*, *12*, 485–502. https://doi.org/10.1017/S1366728909990228 (cit. on p. 103)

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 (cit. on p. 55)

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368. https://doi.org/10.1016/0749-596X(86)90006-9 (cit. on pp. 2, 8)

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII* (pp. 559–586). Lawrence Erlbaum Associates. (Cit. on pp. 2, 8).

Frazier, L., & Clifton, C. (1996). *Construal*. MIT Press. (Cit. on pp. 2, 8).

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies* (Doctoral dissertation). University of Massachusetts. Amherst, MA. (Cit. on pp. 2, 8).

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210. https://doi.org/10.1016/0010-0285(82)90008-1 (cit. on pp. 2, 3, 7–9)

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, *40*(3), 575–586. https://pubmed.ncbi.nlm.nih.gov/6518241/ (cit. on p. 77)

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC. (Cit. on pp. 21, 22, 69–71).

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 (cit. on pp. 5, 53)

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 (cit. on pp. 70, 72)

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). MIT Press. (Cit. on p. 47).

Glaser, Y. G., Martin, R. C., Van Dyke, J. A., Hamilton, A. C., & Tan, Y. (2013). Neural basis of semantic and syntactic interference in sentence comprehension. *Brain and language*, *126*(3), 314–326. https://doi.org/10.1016/j.bandl.2013.06.006 (cit. on pp. 10, 48)

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411–1423. https://doi.org/10.1037/0278-7393.27.6.1411 (cit. on pp. 2, 48)

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, *51*(1), 97–114. https://doi.org/10.1016/j.jml.2004.02.003 (cit. on pp. 2, 10, 48)

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of experimental psychology: Learning, memory, and cognition*, *32*(6), 1304–1321. https://doi.org/10.1037/0278-7393.32.6.1304 (cit. on pp. 2, 11, 48)

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 (cit. on pp. 2, 48–50, 108)

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 (cit. on p. 73)

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: An R Package for Estimating Normalizing Constants. (Cit. on p. 73).

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 (cit. on pp. 54, 55)

Hobbs, B. P., & Carlin, B. P. (2008). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(1), 54–80. https://doi.org/10.1080/10543400701668266 (cit. on p. 77)

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 (cit. on pp. 54, 103, 104, 110)

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 (cit. on pp. 15, 34, 48, 52, 53, 85, 102, 108, 110)

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*, 104063. https://doi.org/10.1016/j.jml.2019.104063 (cit. on pp. 2, 19, 48)

Jeffreys, H. (1961). *Theory of probability*. Clarendon Press. (Cit. on pp. 71, 72).

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149. https://doi.org/10.1037/0033-295X.99.1.122 (cit. on pp. 3, 9, 49)

Kamide, Y., & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes*, *14*(5–6), 631–662. https://doi.org/10.1080/016909699386211 (cit. on p. 40)

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 (cit. on p. 71)

Kathryn Bock, J., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67. https://doi.org/10.1016/0010-0277(85)90023-X (cit. on p. 39)

Keshev, M., & Meltzer-Asscher, A. (2019). Distant relatives: Resumptive pronouns can inherit agreement features of implied antecedents. *32nd Annual CUNY Conference on Human Sentence Processing* (cit. on p. 48).

King, J., Andrews, C., & Wagers, M. (2012). Do reflexives always find a grammatical antecedent for themselves? *25th Annual CUNY Conference on Human Sentence Processing* (cit. on p. 103).

Koesterich, N., Keshev, M., Shamai, D., & Meltzer-Asscher, A. (2021). Interference in the comprehension of filler-gap and filler-resumptive dependencies. *34th Annual CUNY Conference on Human Sentence Processing* (cit. on pp. 38, 48).

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*, 627–645. https://doi.org/10.1023/A:1026528912821 (cit. on p. 40)

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press. (Cit. on pp. 21, 63, 77).

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 (cit. on p. 82)

Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing* (Doctoral dissertation). University of Maryland. College Park, MD. (Cit. on p. 36).

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 (cit. on pp. 10, 33, 54)

Lange, E. B., & Oberauer, K. (2005). Overwriting of phonemic features in serial recall. *Memory*, *13*(3-4), 333–339. https://doi.org/10.1080/09658210344000378 (cit. on p. 38)

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 (cit. on p. 54)

Laurinavichyute, A., & von der Malsburg, T. (2021a). Agreement attraction in grammatical sentences arises only in the good-enough processing mode. *34th Annual CUNY Conference on Human Sentence Processing* (cit. on p. 84).

Laurinavichyute, A., & von der Malsburg, T. (2021b). *Semantic attraction in sentence comprehension*. https://doi.org/10.31234/osf.io/hk9nc. (Cit. on pp. 11, 34, 48)

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759. (Cit. on pp. 71, 72)

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 (cit. on pp. 22, 70)

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 (cit. on pp. iii, 3, 50)

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93–115 (cit. on p. 15).

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 University Press. (Cit. on p. 47).

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 (cit. on pp. iii, 1, 7, 9, 19, 38, 47, 48)

Logačev, P., & Vasishth, S. (2013). *Em2: A package for computing reading time measures for psycholinguistics* [R package version 0.9]. https://cran.r-project.org/src/contrib/Archive/em2/. (Cit. on pp. 21, 74)

Logačev, P., & Vasishth, S. (2016). 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 (cit. on pp. iv, 4, 56, 84)

Lowder, M., & Gordon, P. C. (2014). Effects of animacy and noun-phrase relatedness on the processing of complex sentences. *Memory & Cognition*, *42*, 794–805. https://doi.org/10.3758/s13421-013-0393-7 (cit. on pp. 2, 11, 34, 48)

MacDonald, M. C., Perlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037/0033-295X.101.4.676 (cit. on pp. 3, 8)

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & 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 (cit. on p. 73)

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 (cit. on pp. 1, 9, 47, 48)

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, *38*, 283–312. https://doi.org/10.1006/jmla.1997.2543 (cit. on pp. 3, 8)

Miller, G. A. (1962). Some psychological studies of grammar. *American Psychologist*, *17*, 748–762. https://doi.org/10.1037/h0044708 (cit. on p. 58)

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). John Wiley & Sons. (Cit. on p. 58).

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 (cit. on p. 72)

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269. https://doi.org/10.3758/BF03213879 (cit. on p. 38)

Neath, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin and Review*, *7*, 403–423. https://doi.org/10.3758/BF03214356 (cit. on p. 38)

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 (cit. on pp. 2, 54)

Nicenboim, B., Schad, D. J., & Vasishth, S. (2021). An Introduction to Bayesian Data Analysis for Cognitive Science. https://vasishth.github.io/bayescogsci/book/. (Cit. on pp. 22, 71)

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, *42*, 1075–1100. https://doi.org/10.1111/cogs.12589 (cit. on pp. 2, 15, 48, 85)

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 (cit. on pp. 54, 72)

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.). John Benjamins Publishing. (Cit. on pp. 54, 55).

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of psycholinguistic research*, *18*(1), 5–19. https://doi.org/10.1007/BF01069043 (cit. on p. 10)

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 (cit. on p. 54)

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife*, *6*, e23383. https://doi.org/10.7554/eLife.23383 (cit. on p. 4)

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*, 601–626. https://doi.org/10.1016/j.jml.2006.08.009 (cit. on pp. iii, 3, 38, 50)

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyper-active gap filling. *Frontiers in Psychology*, *6*, 384. https://doi.org/10.3389/fpsyg.2015.00384 (cit. on p. 40)

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 (cit. on p. 37)

Paape, D., Hemforth, B., & Vasishth, S. (2018). Processing of ellipsis with garden-path antecedents in French and German: Evidence from eye tracking. *PLOS ONE*, *13*(6), 1–46. https://doi.org/10.1371/journal.pone.0198620 (cit. on pp. 32, 37)

Parker, D. (2019). Cue combinatorics in memory retrieval for anaphora. *Cognitive Science*, *43*(3), e12715. https://doi.org/10.1111/cogs.12715 (cit. on p. 10)

Parker, D., & Phillips, C. (2014). Selective priority for structure in memory retrieval. *27th Annual CUNY Conference on Human Sentence Processing* (cit. on pp. 104, 110).

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 (cit. on pp. 2, 15, 48, 103, 104)

Patil, U., Vasishth, S., & Lewis, R. L. (2016). Retrieval interference in syntactic processing: The case of reflexive binding in English [Special Issue on Encoding and Navigating Linguistic Representations in Memory]. *Frontiers* *in Psychology*, *7*, 329. https://doi.org/10.3389/fpsyg.2016.00329 (cit. on p. 103)

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*, 427–456. https://doi.org/10.1006/jmla.1999.2653 (cit. on pp. 33, 103, 104)

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 940–961. https://doi.org/10.1037/0278-7393.24.4.940 (cit. on p. 9)

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 (cit. on p. 73)

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/. (Cit. on p. 21)

Rabe, 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 (cit. on p. 80)

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 (cit. on pp. 21, 74)

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 358–374. https://doi.org/10.1016/S0022-5371(83)90236-0 (cit. on p. 9)

Rich, S., & Wagers, M. (2020). Semantic similarity and temporal contiguity in subject-verb dependency processing. *33rd Annual CUNY Conference on Human Sentence Processing* (cit. on pp. 11, 48).

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 (cit. on p. 72)

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). *Workflow Techniques for the Robust Use of Bayes Factors*. https://arxiv.org/abs/2103.08744v2. (Cit. on p. 71)

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 (cit. on pp. 2, 48)

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630 (cit. on p. 86)

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 (cit. on p. 72)

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 (cit. on p. 50)

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, *124*, 101356. https://doi.org/10.1016/j.cogpsych.2020.101356 (cit. on p. 38)

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 357–416. https://doi.org/10.2307/2983527 (cit. on p. 77)

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542–562. https://doi.org/10.1016/S0749-596X(02)00536-3 (cit. on pp. iii, 10, 33, 36, 54, 104, 110)

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 (cit. on pp. iv, 4, 56, 57, 84)

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370. https://doi.org/10.1016/j.jml.2004.01.001 (cit. on pp. 9, 11, 17, 34)

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 431–450. https://doi.org/10.1037/0278-7393.30.2.431 (cit. on pp. 3, 8)

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 (cit. on p. 48)

Traxler, M. J. (2002). Plausibility and subcategorization preference in children's processing of temporarily ambiguous sentences: Evidence from self-paced reading. *The Quarterly Journal of Experimental Psychology Section A*, *55*(1), 75–96. https://doi.org/10.1080/02724980143000172 (cit. on p. 9)

Traxler, M. J. (2005). Plausibility and verb subcategorization in temporarily ambiguous sentences: Evidence from self-paced reading. *Journal of Psycholinguistic Research*, *34*, 1–30. https://doi.org/10.1007/s10936-005-3629-2 (cit. on p. 9)

Traxler, M. J., & Frazier, L. (2008). The role of pragmatic principles in resolving attachment ambiguities: Evidence from eye movements. *Memory & Cognition*, *36*, 314–328. https://doi.org/10.3758/MC.36.2.314 (cit. on pp. 3, 9)

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592. https://doi.org/10.1006/jmla.1998.2600 (cit. on p. 56)

Trueswell, J. C., Tanenhaus, M., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285–318. https://doi.org/10.1006/jmla.1994.1014 (cit. on pp. 3, 9)

Trueswell, J. C., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden paths. *19*, 528–553. https://doi.org/10.1037/0278-7393.19.3.528 (cit. on pp. 3, 7–9)

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 (cit. on pp. ii, iii, 1–3, 7, 9–16, 21, 23, 24, 27, 33–35, 37, 38, 47, 48, 52, 54, 102–105, 111)

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 (cit. on pp. iii, 3, 48, 51, 52, 105, 108, 113)

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 (cit. on pp. ii, 1, 2, 9, 10, 34, 35, 37, 47, 48, 52, 54, 102, 104, 105)

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 (cit. on pp. iii, 2–5, 46, 48, 50–54, 57, 58, 66, 73, 77, 81–83, 86, 95, 105, 108, 111–113, 115)

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 (cit. on pp. ii, iii, 1–3, 7, 9–12, 15, 34–36, 47, 48, 52, 54, 85, 102, 104, 105, 110, 113)

Van Gompel, R. P. G., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *52*(2), 284–307. https://doi.org/10.1016/j.jml.2004.11.003 (cit. on p. 56)

Van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, *45*(2), 225–258. https://doi.org/10.1006/jmla.2001.2773 (cit. on p. 56)

Van Gompel, R., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 621–684). North-Holland. https://doi.org/10.1016/B978-008043642-5/50029-2. (Cit. on p. 33)

Vasishth, S. (2019). Power plot [Available under a CC-BY4.0 license]. https://doi.org/https://doi.org/10.6084/m9.figshare.10281905.v1. (Cit. on p. 53)

Vasishth, S. (2020). Using approximate Bayesian computation for estimating parameters in the cue-based retrieval model of sentence processing. *MethodsX*, *7*, 100850. https://doi.org/https://doi.org/10.1016/j.mex.2020.100850 (cit. on p. 19)

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794. https://doi.org/10.1353/LAN.2006.0236 (cit. on pp. 36, 54)

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 (cit. on pp. 5, 22, 37, 53, 54, 85)

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134. https://doi.org/10.1002/wcs.1209 (cit. on p. 33)

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology*, *9*, 2. https://doi.org/10.3389/fpsyg.2018.00002 (cit. on pp. 37, 38)

Wagers, M. (2008). *The structure of memory meets memory for structure in linguistic cognition* (Doctoral dissertation). University of Maryland, College Park, MD, USA. (Cit. on pp. 14, 37, 38).

Wagers, M., 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 (cit. on pp. 2, 10, 48, 86, 113)

Wagers, M., & McElree, B. (2009). Focal attention and the timing of memory retrieval in language comprehension. *Architectures and Mechanisms for Language Processing Conference* (cit. on p. 40).

Ward, P., & Sturt, P. (2007). Linguistic focus and memory: An eye movement study. *Memory & Cognition*, *35*, 73–86. https://doi.org/10.3758/bf03195944 (cit. on p. 58)

Yadav, H., Paape, D., Smith, G., Dillon, B., & Vasishth, S. (2021). *Individual differences in cue-weighting in sentence comprehension: An evaluation using Approximate Bayesian Computation*. https://doi.org/10.31234/osf.io/4jdu5. (Cit. on p. 10)

Yadav, H., Smith, G., & Vasishth, S. (2021). Feature encoding modulates cue-based retrieval: Modeling interference effects in both grammatical and ungrammatical sentences. *Proceedings of the Cognitive Science conference*. https://psyarxiv.com/76aex/ (cit. on p. 38)

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685 (cit. on p. 86)