GLOSSA psycholinguistics

Winkowski, J., Nouwen, R., & Dotlačil, J. (2024). Searching for semantic distance effects. *Glossa Psycholinguistics*, 3(1): 18, pp. 1–44. DOI: https://doi.org/10.5070/G60111681

eScholarship University of California

Searching for semantic distance effects

Jan Winkowski, Institute for Language Sciences, Utrecht University, NL, j.l.winkowski@uu.nl
Rick Nouwen, Institute for Language Sciences, Utrecht University, NL, r.w.f.nouwen@uu.nl
Jakub Dotlačil, Institute for Language Sciences, Utrecht University, NL, j.dotlacil@uu.nl

Language processing relies on memory. There exists a considerable body of literature on retrieval in *sentence processing* and, in particular, on cases involving recall of syntax-relevant information. There is no reason to doubt, however, that memory is involved in semantic aspects of language processing as well. In this work, we look at the case of additive presuppositions, such as those involved in interpreting the additive particle *too*. When one hears *Mary went to the party, too*, one should recall that someone other than Mary went to the party. We make the case that, as a starting hypothesis, it would be expected that the retrieval of this kind of information should share basic features of memory processes in language with the better-known cases of recall involved in syntactic parsing. In particular, we argue that, given certain assumptions and linking hypotheses, all prominent retrieval theories predict the existence of *distance effects* for the recall of previous information, independent of whether the recall is driven by syntactic or semantic, sentential or inter-sentential, considerations. As the distance increases, so does the difficulty of processing. We test this prediction in four experiments that investigate the role of retrieval in interpreting *too*. Using the Bayesian hierarchical modelling paradigm, we find evidence in two self-paced reading experiments that it takes more time to read sentences with *too* when the distance between the trigger and its antecedent is greater, compared to a baseline that lacks the presupposition trigger. This result shows that theories of the role of memory in language are just as relevant to the domain of discourse interpretation as they are to syntactic parsing. The fact that our evidence was relatively hard to find, however, suggests additionally that there are interesting differences to explore in the future.

1. Introduction

In linguistics, a lot of work has focused on studying the properties of syntactic dependencies (Chomsky, 1957; Tesnière, 1965). Research into syntactic dependencies has been accompanied by experimental research in psycholinguistics on the nature and properties of memory systems deployed in the resolution of syntactic dependencies (Jäger et al., 2017; Lewis & Vasishth, 2005; McElree et al., 2003; Van Dyke & McElree, 2011; Wagers et al., 2009).

We believe that the reliance on memory does not stop at sentence boundaries. Language users have to remember elements and information spread out through whole discourses to arrive at the correct interpretation of complete texts. In linguistics and philosophy, the properties of discourse dependencies have been investigated in a formally explicit way by branches of formal semantics, such as discourse semantics and dynamic semantics (Asher & Lascarides, 2003; Heim, 1982; Kamp, 1981; Kamp & Reyle, 1993; Nouwen et al., 2016). While this research is well established, it has not been supplemented by detailed psycholinguistic investigations into the properties of memory systems used for the resolution of such discourse dependencies to the same extent that properties of memory processes for syntactic dependencies have been studied.

In this article, we investigate the memory processes involved in resolving one particular case of discourse dependencies. We focus on the particle *too*, which signals a dependency that can stretch back many sentences to be resolved (Kripke, 2009). Consider the following example:

(1) A contestant started opening boxes one by one. The first box was empty. The second box contained a piece of gold. The contestant hesitated before opening the third box. This box was empty, too.

The particle *too* triggers the presupposition that there was another box apart from *this box* that was empty (for instance, van der Sandt & Geurts, 2001). Resolving this presupposition is only feasible if readers recall what was mentioned at the beginning of the discourse. Such dependencies, spanning potentially large portions of a discourse, are subject to various structural and non-structural constraints, investigated in the field of semantics (Kripke, 2009; Szabolcsi, 2017; van der Sandt, 1992). Our goal in this article is to investigate how the resolution of this presupposition is affected by the distance between *too* and its antecedent, the information present in the sentence *The first box was empty* in (1). The investigation allows us to test a basic prediction made if we assume that the processes involved in recalling semantic information, potentially across sentence boundaries, share basic features with syntactic sentence processing.

The article is designed in the following way. In the next, section we discuss distance effects that have been found to exist for syntactic retrieval, and we show that, given certain assumptions, we can predict similar effects for the semantic processing of additives like *too*. Sections 3, 4, 5, and 6 present two acceptability judgment experiments and two self-paced reading experiments. Then, in Section 7, we use the Bayesian paradigm to inspect the pooled data from all experiments.

The model fit to the pooled data will present evidence that confirms our prediction. Finally, we discuss some important questions raised by these findings.

2. Distance effects in syntactic and semantic processing

2.1 Distance effects in syntactic dependencies

Distance effects (also called *locality effects*) occur when processing complexity is affected by the distance between words that are involved in some kind of dependency. An example of such a dependency is the relation between the head noun and the verb of a relative clause, as *the administrator – supervised* in (2) (based on Grodner & Gibson, 2005).

(2) a. The administrator who the nurse supervised scolded the medic.
b. The administrator who the nurse who was from the clinic supervised scolded the medic.

The additional material in (2b), compared to (2a), results in processing difficulties. The difficulty manifests itself as increased reading times in eye-tracking while reading studies and self-paced reading studies (Bartek et al., 2011; Grodner & Gibson, 2005; Nicenboim et al., 2016, among others) or as a decrease in accuracy in studies using the speed-accuracy tradeoff paradigm (e.g. McElree, 2000).

Assuming a certain view of how memory is involved in language processing (Anderson, 1991; Lewis et al., 2006), there are two possible sources for such findings: decay and interference (see also Bartek et al., 2011). Both of these can play a role in processing. As the distance to the retrieval target grows, the activation of that target decays. At the same time, with greater distance comes an increase in the number of similar elements that compete with the antecedent for retrieval. This competition reduces accuracy, but can also have an impact on speed when faulty resolutions need to be rectified.

Distance effects are accounted for in different ways by various theoretical frameworks. For instance, in dependency locality theory (Gibson, 2000), the integration cost of connecting some current structure to a previous structure depends directly on the number of discourse entities (discourse referents, events, times) that have been introduced in the meantime.

Another dominant theory of how memory is involved in sentence processing posits that dependents in long-distance dependencies are retrieved via a so-called cue-based mechanism (Jäger et al., 2017; Lewis & Vasishth, 2005; McElree, 2001; McElree et al., 2003; Nicenboim et al., 2015; Nicenboim & Vasishth, 2018, and others). This means that when an item previously seen is needed to parse a dependency, it is retrieved from memory using cues determined by the word being processed. Consider (3) (taken from Van Dyke, 2007).

(3) The worker was surprised that the resident who was living near the dangerous neighbor **was complaining** about the investigation.

When reading the phrase *was complaining*, the parser has to find the subject of the clause (*the resident*) to arrive at the correct interpretation. To retrieve the subject from memory, the parser uses cues. In this way, the search is restricted to the elements that satisfy the constraints introduced by the verb phrase. We can assume that the parser would consider the cues {*noun phrase*}, {*local subject*} and {*animate*}, among others.

In examples like (3), there are several elements that match the cue {*animate*} and {*noun phrase*} without matching the cue {*local subject*}: *the worker*, and *the neighbour* are such cases. Such partially-matching elements are commonly called *distractors*. The interaction of phrases carrying partially matching sets of features with the retrieval process is called *cue interference* and is usually taken as evidence in favour of the cue-based retrieval mechanism (Jäger et al., 2017). Due to cue interference, distractors can in some environments cause misretrieval and slowdown in reading times at the moment of dependency resolution (e.g. Jäger et al., 2017; Nicenboim & Vasishth, 2018).

There is a plethora of research studying the cue-based retrieval mechanism in the syntactic domain (see Jäger et al., 2017, for an overview and a metaanalysis). This body of work extends also to syntactic phenomena involving retrieval of semantic information, such as the licensing of negative polarity items (Parker & Phillips, 2016) and to the role of semantic information in resolving syntactic dependencies (e.g. Cunnings & Sturt, 2018; Laurinavichyute & von der Malsburg, 2022).

In contrast to this research, the investigation of retrieval mechanisms for discourse dependencies is much more limited. Most studies in psycholinguistics that go beyond the domain of syntax consider pronoun resolution (Kush & Eik, 2019; Kush & Dillon, 2021; Schmitz et al., 2024). Yet, it is clear that there are other dependencies in discourse that could be investigated from this perspective.

In what follows, we will show in a theory-independent way that the processing of *too* relies on memory. This fact will allow us to derive experimental predictions using cue-based theories of memory retrieval. We note here that the predictions we are to discuss and test are likely shared with various other theories of memory use in language comprehension, like dependency locality theory. Our goal here is not to decide which of the theories is right. Rather, we want to show that under quite general conditions on memory, which we will make explicit using the cue-based retrieval model as our illustrative framework, we can establish whether there is a cost to processing *too* due to retrieval.

2.2 Processing too

In contrast to studies concerned with syntactic dependencies, the phenomenon we are interested in is largely independent of syntax. Yet, the additive presuppositions we focus on in what follows *do* clearly qualify as *dependencies*. So, in this sense, they *are* quite close to what has been studied

in the syntactic domain (see also Brasoveanu & Dotlačil, 2020; Chen & Husband, 2018, for a similar argument).

To see this, consider a sentence like (4). This triggers a presupposition that is connected to the so-called *associate* of the additive *too*, here assumed to be *Sue*: namely, that someone other than Sue lives in France.

(4) Sue lives in France, too.

Given this presupposition, we can now demonstrate that additives like *too* are involved in dependencies and that processing them relies on retrieval. *Too* is subject to a so-called *strong contextual felicity constraint* (Tonhauser et al., 2013), which means that it is infelicitous whenever the presupposition is not satisfied in the preceding context.[1] Consider the example in (5) (intended as an utterance out of the blue).

(5) Anne lives in Brussels. Sue lives in France, #too.

The fact that we can recognise that (5) is infelicitous means that we use our knowledge of the preceding context when we process the additive. To be more specific, our capacity to judge the second sentence in (5) infelicitous after encountering *too* entails that upon encountering the additive, a retrieval process is triggered to match the presupposition. The result of this process is known to have an effect on sentence processing. Schwarz (2007) shows that reading times are longer in conditions where a trigger lacks an antecedent, compared to a condition where an antecedent is available. Presumably, such effects, as well as the infelicity judgment of (5), are caused by a failure of the match following the retrieval process.

The basis for this retrieval process is information-structural rather than syntactic. There is a general consensus that the associate of the additive is in focus (e.g. Geurts & van der Sandt, 2004; Ruys, 2015) or is a contrastive topic (e.g. Krifka, 1998). The semantic import of bearing focus (and being a contrastive topic) is the activation of alternatives (Krifka, 2008; Rooth, 1992). Such alternatives usually trigger an exhaustification implicature, the inference that all alternatives except for the asserted one are false (Fox & Katzir, 2011; Kiss, 1998; Rooth, 1992). For instance, *Sue lives in France* (with the subject bearing focus) triggers an implicature that none of the other contextually relevant individuals (i.e. alternatives to *Sue*) live in France. The additive in (4) overtly cancels that implicature, by presupposing that one of the alternatives *is* true (Bade, 2014; Krifka, 1998; Sæbø, 2004).

[1] There are some well-known cases where additives *are* felicitous out of the blue. For instance: *The rich should suffer, too!* Such examples, in which the presupposition of *too* is accommodated, play an important role in the debate on the nature of additive presupposition. In the remainder of the article, however, we only look at examples where the presupposition is not accommodated. In other words, in all the cases we look at, it is evident that some memory process is involved.

The additive's presupposition can be obtained by looking at the alternative propositions triggered by the associate bearing focus. For instance, focus on *Sue* in the second sentence of (5) creates alternatives like *Peter lives in France, Bob lives in France, Anne lives in France*. The presupposition of *too* says that at least one such alternative is true. In other words, if the presupposition is satisfied textually, then the preceding discourse should entail one of these alternatives. Since *Anne lives in Paris* entails *Anne lives in France*, the presupposition is met in an example like (6), but since *Anne lives in Brussels* does not have that entailment, (5) ends up being infelicitous.

(6) Anne lives in Paris. Sue lives in France, too.

Given this picture, any proposition entailed in the preceding discourse is a potential antecedent to the additive. Importantly, this means that the number of potential antecedents increases as the preceding discourse gets larger. This is not dependent on the number of preceding clauses, but rather on the number of constituents. Consider, for instance, (7):

(7) Last week, Sue gave a 14 year old bottle of wine to John. Sue's cat is 14 years old, too. But that's a coincidence.

Examples like this show that the addition of any constituent has the potential to be responsible for an entailment that could satisfy some presupposition down the line. Focus on the subject of the second sentence in (7) activates alternatives of the form *X is 14 years old*. The additive triggers the presupposition that one of these alternatives is true, where *X* is not Sue's cat. The first sentence entails that the bottle of wine given to John is 14 years old and, as a consequence, the presupposition is satisfied. However, obviously, without the addition of the AP *14 year old* modifying *bottle of wine*, the presupposition would not be met and the discourse in (7) would be infelicitous. This observation has consequences for how we may understand the retrieval process involved in additive presuppositional dependencies. As examples like (7) show, the number of potential dependents (and, so, the number of distractors) increases not just with the number of intervening clauses, but also with the number of intervening constituents.

There are obvious parallels between the retrieval process involved in syntactic dependencies and that involved in presuppositions. Just as the number of syntactic attachment points increases as the sentence proceeds, there is a comparable accumulation of entailments as the discourse progresses. Consequently, just as we expect processing complexity to increase with increased distance within a syntactic dependency, all else being equal, we expect a similar effect of complexity to occur with presuppositions whenever the antecedent entailment is triggered earlier in the discourse. Put simply, if we assume that distance effects are due to inherent constraints on the memory processes involved in language processing, and given that additive presuppositions involve retrieval, we come to expect that the processing of additive presuppositions will display distance effects.

It should be noted, though, that it is possible to satisfy an additive presupposition without identifying a single word or phrase responsible for this entailment. Take the following example:

(8) By Thanksgiving, his spots disappeared, his coat turned a rich, dull, field-mouse brown. He gained weight, abandoned the goat's milk for forage. Buck became a fixture on Church Road, bounding alongside Skye when she went jogging, greeting vehicles at the mailbox, startling the FedEx delivery man, and charming neighbors into leaving treats out for him — a few apples, some oatmeal, a pile of hickory nuts. Billy was changing, too. (taken from Corpus of Contemporary American English, https://www.english-corpora.org/coca/)

Clearly the presupposition is satisfied here; the set of sentences in the preceding discourse seems to entail the proposition *Buck was changing* for some relevant meaning of *changing*. However, the word itself does not appear in that discourse. So, in contexts like this, it would be hard to pinpoint a location in the preceding text where the entailment originates. This is not surprising; since additive presuppositions involve dependencies built on entailments, not all such dependencies will be reducible to a dependency between *too* and a word or a phrase.[2] We purposefully avoid examples like this in our stimuli, which will allow us to match the antecedent of the presuppositional dependency (an entailment) with a position in the preceding text. This will allow us to test the expectation that presuppositional dependencies are subject to distance effects.

This expectation is dependent on presuppositions being triggered rapidly. If presuppositions are slow inferences, the comparison to syntactic processing becomes moot. In the literature, presupposition processing speed has been connected to the question whether presuppositions are more like semantic entailment (Heim, 1982) or more like pragmatic implicature (e.g. Schlenker, 2008). This is because, for a while, it was thought that pragmatic inferences are typically slow and costly (e.g. Huang & Snedeker, 2011, views tend to be more nuanced now, e.g. Degen & Tanenhaus, 2016). It has proven to be difficult to use acceptance/rejection reaction time studies to compare presupposition and implicature (Chemla & Bott, 2013; Romoli & Schwarz, 2015), but there is clear evidence from other paradigms that presuppositions are not delayed. Using the visual world paradigm, Romoli et al. (2015); Schwarz (2015) show that fixation shifts to pictures representing potential antecedents of additive presupposition triggers very soon after such a trigger is encountered. This suggests that the retrieval process is started rapidly. That is, antecedent retrieval is like the rapid processing of syntactic information. As a result, comparison with the resolution of syntactic dependencies makes sense from a processing point of view.

[2] This does not mean, however, that the notion of distance is not applicable to the general case. Even in cases such as (8), one can identify a point at which enough information has been accumulated for the relevant entailment to be satisfied. Precise identification might be difficult, because the presupposition triggered by *too* does not need an exact match to be resolved, but it is possible to identify the last contributing expression. E.g. in (8), adding material after *hickory nuts* will lengthen the distance between the antecedent and the trigger.

At the same time, there are also obvious differences between retrieval of syntactic material and retrieval involving semantics. It is much less evident what kind of cues could guide the retrieval of entailments. Entailments do not always coincide with linguistic events, such as the processing of a word or constituent. In this sense, it is harder to translate them into cues than, for instance, connecting a phrase to the feature {*subject*}. Still, such connections *are* proposed in representational theories of (discourse) semantics, most prominently in Discourse Representation Theory (DRT; Kamp, 1981; Kamp & Reyle, 1993). In DRT, an incremental semantic representation of the discourse is formed as the discourse progresses. The ensuing representations impose constraints on what information is accessible where in the discourse. As such, it offers a representational basis for semantic notions like entailment and, therefore, it could potentially act as a framework for imagining what a cue-based mechanism for presupposition retrieval looks like. In fact, DRT has been an influential framework for theories of presupposition (e.g. Geurts, 1999; van der Sandt, 1992).

2.3 Predictions

The items used in the experiments below allow us to simplify, to some extent, the complex picture introduced above. Consider, for instance, (9), which is a stimulus from Experiment 1.

(9) The cook is a dancer and the waiter, who is a great boxer from southern Amsterdam, dances too, I have been told recently.

When encountering the verb *dances* followed by the additive *too*, the parser needs to find the information that makes the additive licensed. In particular, some part of the discourse (representation) needs to be found that entails *X dances* for something other than $X = $ *the waiter*. The fully matching part of the discourse is *The cook is a dancer*, since this clause entails that the cook dances, but between that and the additive there are many constituents that introduce predicate-argument propositions that fail to match the presupposition. For instance, *a great boxer* has the same subject as *dances* and is thus unsuitable. The same goes for *from southern Amsterdam*. Both these predicates also fail to subsume the property of *dancing*, which is another reason why they are unsuitable antecedents. Still, they do qualify as distractors, since they introduce entailments (that the waiter is a great boxer, that the waiter is from southern Amsterdam) that potentially could satisfy upcoming presuppositions. This means we could simplify things by matching distractors to *predicates* appearing somewhere else in the text. Such predicates can be realised as a VP, an appositive noun, a PP, etc.[3] What matters is the semantic content present in the discourse. We assume that {*predicate*} can be used as a cue when searching for a relevant

[3] The connection between phrases and predicates is not always trivial. For our items, it is enough to know at which point the predicate is established in discourse, so that it can serve as an antecedent. This also means that the same point in the discourse can establish multiple predicates, for instance, in cases where a transitive verb phrase contains a noun denoting a predicate.

item to retrieve. Since predicates are ubiquitous in language, we believe that this makes it theoretically possible to investigate whether the cue-based retrieval mechanism is a plausible theory of the on-line interpretation of additives. Under this theory, we expect interference due to the number of predicates, which can be approximated by the number of words intervening between the antecedent and the additive. Based on this assumption, the cue-based retrieval mechanism is compatible with the observation of more errors and/or slower reading times as the number of words intervening between the antecedent and the additive is increased.

Let us spell out the prediction in more detail. Assuming that predicates act as distractors, they interfere with the retrieval process. The distractors could be retrieved erroneously. If readers need to correct the misretrieval by attempting retrieval again, or they notice that the information retrieved does not match what they wanted to retrieve and end up with an infelicitous discourse, an increase in their reading time is expected (Nicenboim & Vasishth, 2018). Alternatively, according to some theories of cue-based retrieval, the presence of distractors should lower the activation of the antecedent, which is consequently harder to retrieve, so its retrieval should take more time (Jäger et al., 2017; Lewis & Vasishth, 2005). Either assumption would derive a slowdown when more words intervene between the antecedent and *too*. However, even outside of the cue-based retrieval mechanism, we predict an interfering role for intervening words. Under models of memory in which memory traces decay over time (Brown et al., 2007, base-level activation; Anderson & Lebiere, 1998), a similar prediction is made. Since increasing the number of words between the antecedent and the additive should increase the time elapsed, it is expected that the antecedent should be harder to retrieve when more words intervene.

In summary, given the assumptions about the various ways retrieval processes work in language processing and given that presupposition resolution involves retrieval, just as syntactic dependency resolution does, we predict that the greater the distance between the trigger and the antecedent, the more difficult it will be to retrieve the antecedent. This difficulty should manifest itself in longer reading times observed on the relevant regions. Besides reading times, we also examine whether the predicted difficulty will manifest itself as diminished acceptability of sentences with greater distance between the trigger and the antecedent.

3. Experiment 1: Acceptability judgment task

3.1 Design

The first experiment was designed to test whether linguistic distance has an influence on the perceived acceptability of sentences with presupposition carried by an additive *too*.

It is usually assumed that processing difficulty has an influence on acceptability (Miller & Chomsky, 1963, Staum Casasanto, Hofmeister, & Sag, 2010 and many others). The relation between the two is not so clear when it comes to distance (locality) effects. For example,

Konieczny (2000) reports conflicting results from acceptability judgments and self-paced reading experiments on locality effects. This suggests that, in this particular domain, processing difficulty and acceptability are not connected in a straightforward way. We decided to collect acceptability judgment data as a sanity check and a first step in investigating the processing of *too*. In case distance effects were observed in the acceptability judgment task, we would have evidence that distance effects play a role even in off-line measures. If we do not observe such effects, the natural next step is to investigate on-line processing of *too*.

We compared sentences like (10a), in which an additive presupposition is triggered by *too*, with sentences like (10b), where there was no such trigger and no presupposition.

(10) a. The cook is a **dancer** and the waiter dances **too**.
b. The cook is a **swimmer** and the waiter dances **often**.

Apart from varying the presence or absence of a presupposition trigger, we also manipulated the length of the sentences used. This varied the distance between the presuppositional antecedent and the trigger, in a way designed to help us observe the differences in acceptability (or reading times) between the trigger/no trigger conditions.

In the experiment, four levels of distance (labelled 0, 4, 8, or 12 for extensions of 0 words, 4–5 words, 8–10 words, or 12–15 words; 4 words change between the conditions at minimum) were crossed with two levels of trigger (*too/nil* – for trigger present or no trigger, respectively). This results in a within-participants design with $4 \times 2 = 8$ conditions. The distance factor was manipulated by increasing the number of words between the trigger site and the antecedent. The trigger factor was manipulated by replacing the trigger with a non-triggering word, usually an adverb, and aligning the sentence meaning with this change. This usually involved changing the nominal predicate to a different noun. An example of an item is shown below.

too

0 The cook is a dancer and the waiter dances too, I have been told recently.

4 The cook is a dancer and the waiter, who is a boxer, dances too, I have been told recently.

8 The cook is a dancer and the waiter, who is a great boxer from southern Amsterdam, dances too, I have been told recently.

12 The cook is a dancer and the waiter, who is a great lightweight boxer from east parts of southern Amsterdam, dances too, I have been told recently.

nil

0 The cook is a swimmer and the waiter dances often, I have been told recently.

4 The cook is a swimmer and the waiter, who is a boxer, dances often, I have been told recently.

8 The cook is a swimmer and the waiter, who is a great boxer from southern Amsterdam, dances often, I have been told recently.

12 The cook is a swimmer and the waiter, who is a great lightweight boxer from east parts of southern Amsterdam, dances often, I have been told recently.

The target items were sentences consisting of two parallel clauses. In the first clause, some property was asserted of the subject using a noun (e.g. *dancer* in (*too*)). In the second clause, this property was recalled using a verb followed by *too* (e.g. *dances too* in (*too*)). In the *nil* condition, the noun in the first clause was unrelated to the verb in the second clause, and *too* was replaced by an adverb (e.g., *swimmer* and *dances often* in (*nil*)).

We constructed 32 items of this kind. The target items were combined with 22 filler items. The fillers were grammatical sentences of similar length to the target items. Additionally, we used 10 items as controls. Their purpose was to check on the engagement of the participants. These were sentences that clearly elicit judgments of unacceptability, because they violate some syntactic or semantic constraint. All the stimuli are available in the supplementary material.

3.2 Participants

32 participants (female = 18 (56%), male = 12 (38%), other/don't want to say = 2 (6%)) were recruited using Prolific (https://www.prolific.co). Participants were compensated 1.97 USD for taking part, which was calculated assuming that the experiment should take around 10 minutes to complete, and given that the effective minimum wage in the United States was 11.80 USD at the time of the experiment ($\frac{1}{6} \times 11.80 = 1.97$). The participants were self-reported English native speakers and self-reported US nationals.

3.3 Procedure

The experiment was conducted using Ibex.[4] The 31 item sets were distributed across 8 lists in a Latin Square design, and participants were assigned at random to each list.[5] Each of the target sentences was interspersed with either a filler item or a control item. The order of items across lists was randomized. The experiment was run from the participants' internet browser.

Before starting the experiment, there was an introduction containing a consent form, a question regarding native language (or languages), and a question about the participant's gender with three possible answers: *male*, *female*, *other/don't want to say*. After the introduction, there was an explanation of the task, followed by a training session with two training items.

During the actual task, participants saw a sentence in the middle of the screen. Below the sentence was a question asking: *How does the sentence sound?* The question was accompanied by a 5-point scale labelled *unnatural* near *1* on the left, and *natural* near *5* on the right. On the top of the screen, a progress bar was displayed. After choosing a score either with a mouse cursor or with a keyboard, the participant saw a screen with the message: *Please wait for the next sentence,* and after one second, they were presented with the next item.

[4] We used Alex Drummond's Ibex farm (https://ibex.spellout.net), which is no longer operational. PCIbex farm (https://farm.pcibex.net/) could be used to re-implement the experiments in their original form.

[5] Item 32 was not used in the actual experiment, because of a coding mistake. We became aware of this after the data was collected.

3.4 Predictions

There are two specific predictions for this experiment. First, we predict that longer items will be judged as worse than the shorter ones. This prediction is in line with the findings of Dillon et al. (2014), who found that longer restrictive relative clauses are judged as worse than shorter ones.

The prediction which is of greater importance to us, concerns the difference in processing difficulty between the sentences with *too* and without *too*. We predict that increasing distance in the sentences with *too* will make them more difficult to process. However, if we test this prediction only on sentences with *too*, we cannot be sure if this difficulty is not just a general effect of the greater distance, independent of *too*. To overcome this difficulty, we are comparing sentences with *too* to sentences without *too*, and we will look for an interaction of the trigger with distance. The difference in difficulty between the long and short sentences should be larger for sentences with *too* than for those without it. We expect this difficulty to be visible as a difference in ratings across relevant conditions. That is, if distance effects in retrieval are visible in an off-line study, we predict a larger decrease in ratings across distance in the *too* conditions than in the *nil* conditions.

3.5 Data analysis

After visual inspection, two participants were removed from the analysis, since their median response to the control sentences was above 3, indicating that they treated purposefully unnatural sentences as quite natural.

A Bayesian hierarchical model was fit to the data obtained. In Bayesian data analysis, one specifies the likelihood, and the prior distributions over the parameters of interest (Gelman et al., 2003; Gelman & Hill, 2006; McElreath, 2018; Nicenboim et al., 2021). The analysis results in posterior probability distributions of plausible values for a given model and data. We report means and 89% credible intervals (CRI), i.e. the range of values for which we can be 89% certain that the true effect lies therein. The mean and the range of credible intervals are supposed to convey the shape of the posterior distribution. We specifically avoid using 95% credible intervals to discourage the readers from conducting unconscious hypothesis tests (see also McElreath, 2018, p. 58).

We used ordinal regression models, which are particularly well suited to the analysis of discrete data (Bürkner & Vuorre, 2019; Veríssimo, 2021). In particular, we used a cumulative model with a logit link function and so-called flexible thresholds (i.e. the estimated distance between the different scores can vary). Models used 4 sampling chains, with 4,000 samples drawn from each chain. Half of these samples were discarded for warm-up; hence each model had 8,000 samples available for the analysis. Trace plots and plots of posterior predictive distribution were visually inspected to identify convergence issues. Additionally, only the models with all $\hat{R} \leq 1.01$, which suggests convergence, were used in the analyses.

All the models used were fit with a full variance-covariance matrix, i.e. they were so-called maximal models. The analysis was conducted in R software for statistical computing (R Core Team, 2021), and, in particular, with the use of the brms package (Bürkner, 2017), which uses the Stan probabilistic language (Stan Development Team, 2021).

The distance, levels of which were labelled with numbers closely matching the number of words between the trigger and its antecedent, was treated as a numeric variable which was scaled and centered before fitting the model. The trigger was sum-contrast coded (*nil* = –1, *too* = 1). Fixed effects of interest were: distance, trigger and their interaction. Participants and items were used as random effects.

We used regularising prior distributions for slopes and intercepts. For the intercept, we used a normal distribution with $\mu = 0$ and $\sigma = 2$. This means that any of the intercepts (i.e. the thresholds for different Likert scale outcomes) should fall with 89% probability between 0.04 and 0.96 on the probability scale. For the slopes, we used a normal distribution with $\mu = 0$ and $\sigma = 0.5$.

This assumes that if we fix the intercept at 0 log-odds (i.e. 0.5 on the probability scale), and the input at 1, then the values of the outcome will be, with 89% probability, between 0.69 and 0.31 on the probability scale. Or, to put it differently, the priors used for the slopes mean that when the intercept is at 0.5 probability, we assume, with 89% probability, that the slope coefficient will change the outcome by at most 0.197 on the probability scale. The priors were additionally examined using prior predictive checks.

Note that the size of the effect depends on the intercept, because of the logit link function. We report the results on the log-odds scale.

The 89% credible intervals for the slopes of interest are summarised in **Figure 1**.

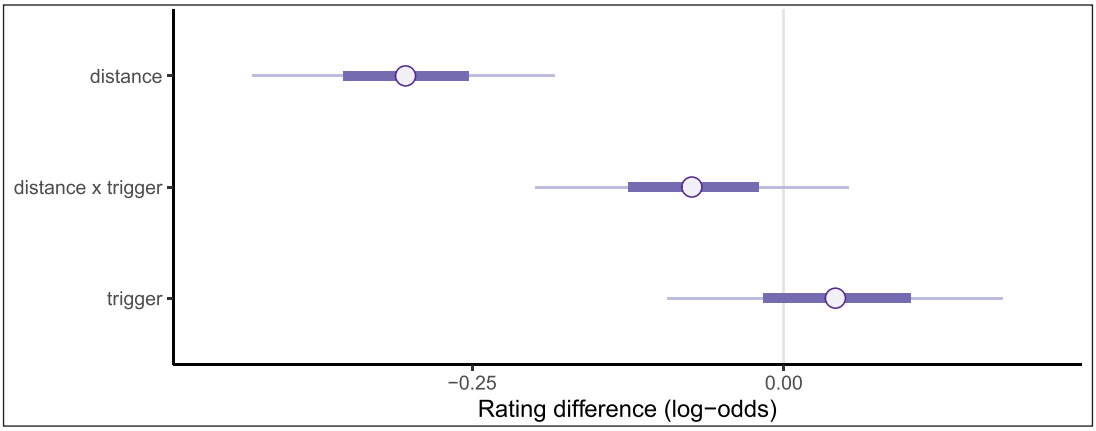

Figure 1: The posterior probability distribution of the parameters of interest in Experiment 1. The thicker lines delimit 50% probability intervals, the thinner lines delimit 89% probability intervals. The point signifies the mean value.

The estimate of the distance parameter indicates that the longer items were judged as less natural (89% CrI = (–0.4, –0.2), mean = –0.3, sd = 0.08).

In the case of the trigger, the sign of the estimate is rather uncertain, with the 89% (–0.09, 0.18) interval lying on the both sides of 0. The mean (= 0.04, sd = 0.09) is positive (i.e. the sentences with *too* were judged as slightly more natural), but the spread of the posterior distribution indicates that the sign of estimate is not conclusive.

For our purposes, the most interesting parameter is the slope of the interaction. The posterior distribution of the slope of the interaction (mean = –0.07, sd = 0.08) has the 89% credible interval (–0.2, 0.05), that is, we can be 89% certain that the value of the interaction lies in this interval. The interval is more negative than positive, but the fact that it spans zero suggests that zero (no effect) is a possible value for the interaction term.

In sum, there is a clear effect of distance (longer items are less natural), but the model does not show that the presence of the trigger and the interaction have an effect on acceptability.

4. Experiment 2: Self-paced reading

4.1 Design

In the self-paced reading task, we used a subset of the conditions and items from the acceptability judgment task. These were chosen based on our intuition, i.e. one of the authors went through all the items and selected the ones they thought sounded good. Some of the remaining items which we thought could be improved were edited for the self-paced reading experiment. The exact changes between the items can be seen at the OSF repository.

With respect to the experimental manipulation, two levels of distance (*short* and *long*, originally 0 and 8) were crossed with two levels of trigger (*too* and *nil*), which resulted in a 2 × 2 within-participant design. We used 24 target items in the self-paced reading task. These were accompanied by 25 filler items.

4.2 Participants

82 participants (female = 44 (54%), male = 37 (45%), other/don't want to say = 1 (1%)) were recruited using Prolific (https://www.prolific.co). They were compensated 3.00 GBP for taking part in the experiment. The experiment took, on average, 12 minutes and 38 seconds to finish; therefore, the participants would get 14.25 GPB per hour, on average. The participants were self-reported English native speakers.

4.3 Procedure

The experiment was conducted in a noncumulative self-paced reading moving window paradigm (Just et al., 1982): during reading, only one word appeared and participants had to

press the space bar to reveal the next word and hide the previous word. In the initial position, every word was masked by a line which was roughly equal to the word's length. When the words disappeared, they went back to the initial position. A comprehension question followed each item.

4.4 Predictions

As was the case for the acceptability judgment experiment, there are also two specific predictions for this experiment. The first prediction follows from the observed relationship between reading speed and distance. It is commonly observed (Demberg & Keller, 2008; Ferreira & Henderson, 1993; and others) that readers speed up with the length of a sentence. Therefore, we predict that the relevant regions in the longer sentences will be read faster than in the shorter sentences.

The prediction which is of greater importance to us concerns the difference in processing difficulty between the sentences with *too* and without *too*. We predict that increasing distance in the sentences with *too* will make them more difficult to process. Following the reasoning from the previous experiment, we should note that if we test this prediction only on sentences with *too*, we cannot distinguish the difficulty due to retrieval from the general effect of longer texts on reading. To overcome this, we are comparing sentences with *too* to sentences without *too*, and we will look for an interaction. The difference in difficulty between the long and short sentences should be larger for sentences with *too* than for those without it. We expect this difficulty to be visible as a difference in reading times across relevant conditions. That is, we predict that compared to the effect of distance in the *nil* conditions, there should be a slowdown effect due to distance in the *too* conditions.

4.5 Data analysis

Before the analysis, we investigated the quality of the responses to the comprehension questions. Out of 82 participants, 1 with the worst ratio ($\leq$ 0.7 correct) of correct responses to the comprehension questions was removed from the results. The median correct response ratio among the remaining participants was equal to 0.9.

Additionally, time measurements which were under 50 ms or above 3,000 ms were removed from the results. Similar data trimming was done, for example, by Futrell et al. (2021). This resulted in the removal of a further 0.2% of the data.

Next, we fitted a Bayesian mixed-effects model to the response time data. This model assumed log-normal likelihood, since reading time data are approximately log-normally distributed and, therefore, our model can resemble the data-generating process more closely (see also Nicenboim et al., 2016, 2018; Rouder et al., 2008; for a more detailed discussion). Besides the choice of priors and likelihood, the details of model fitting were similar to those in

the analysis of the acceptability judgment task. We analyzed the critical region (the word *too* and the corresponding word in the *nil* condition), the post-critical region (the word immediately following the critical region), and the post-post-critical region (two words after the critical region). The experimental factor *trigger* was coded using sum contrasts, with the *nil* condition as –1 and the *too* condition as 1. The experimental factor *distance* was also coded with sum contrasts. The *long* condition was coded as 1 and the *short* condition as –1. Both of these variables were entered into the model as fixed effects. Participants and items were entered as random effects.

Besides models quantifying the effects of interest on the entirety of the data, we also fit so-called nested models to quantify the effect of distance for each of the trigger conditions. These models were implemented by specifying two predictors: too_distance and nil_distance. For too_distance, we coded the conditions where the trigger was *too* and distance was *long* as 1, the conditions where the trigger was *too* and the distance was *short* as –1, and the *nil* conditions as 0. Analogously, for nil_distance, the conditions where the trigger was *nil* and the distance was *long* were coded as 1, the conditions where the trigger was *nil* and the distance was *short* were coded as –1, and the *too* conditions were coded as 0. Both of these predictors were then used to fit varying slopes and varying intercepts models. These models provide a comparison of the effect of distance separately for each trigger condition. They used 4 sampling chains, with 6,000 samples drawn from each chain. 2,000 samples from each chain were discarded for warm-up; hence each model had 16,000 samples available for the analysis.

For the intercept, we used a normal distribution with $\mu = 6$ and $\sigma = 1.5$. This means that the intercept (i.e. the grand mean in our models) is assumed to be between 36.7 ms and 4435 ms, with 89% probability. For the slopes, we used a normal distribution with $\mu = 0$ and $\sigma = 0.1$. This assumes that if we fix the intercept at 6 (which is $e^6 = 403.4$ ms), and the input at 1, the output will be between 473.3 and 343.8 ms, with 89% probability. Note that the degree to which slope coefficients influence the outcome depends on the intercept, because of the log-normal likelihood.

If the prior distribution of the slopes seems too constraining, one should consider that it is in line with the estimates of the meta-analysis of interference effects studies (Jäger et al., 2017). In that meta-analysis, the estimate of the largest relevant effect (p. 328, Table 4: target mismatch configurations, *AND/OR* prominence effect) was 39.3 ms (95% CrI (10.8, 67.9)). One can use the credible interval to guide the selection of the prior distribution for the standard deviation. Following Schad et al. (2021), we come up with an estimate for the σ parameter on the log scale equal to 0.017. Since we cannot be sure whether effects related to the processing of semantic/pragmatic dependencies are of a similar order of magnitude, we settle on a much more uncertain prior of 0.1. Furthermore, we evaluated the prior distributions with prior predictive checks. The visual summaries of the checks are available at the OSF repository. As a prior distribution for both the standard deviations of the random effects, and for the residual standard

deviation, we used a truncated normal with $\mu = 0$ and $\sigma = 1$. For the random effects correlation between the intercept and the slope, we used the LKJ distribution (Lewandowski et al., 2009; Stan Development Team, 2021) with $\eta = 2$.

The results are reported on the log-ms scale.

4.6 Results

The descriptive summaries of the reading times across conditions from one word before to two words after the critical region (the additive *too* and the corresponding word in the *nil* condition) are shown on **Figure 2**.

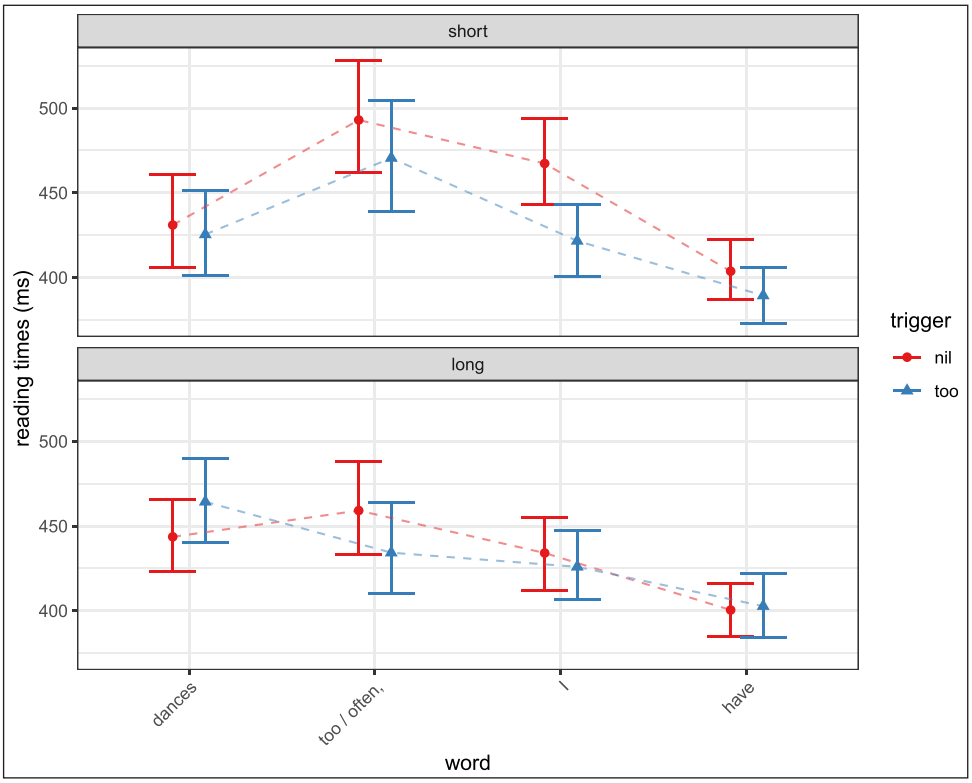

Figure 2: Descriptive summary of reading times in Experiment 2. The points connected by dashed lines denote mean reading times across conditions. These were computed by taking an average of all the data points for each condition, i.e. without grouping by participants or items. The bars represent bootstrapped 0.95 confidence intervals.

Let us turn to the findings of the Bayesian mixed-effects model. First, we discuss the results for the critical region (i.e. *too* or the parallel word), then the post-critical region (i.e. the word after the *too* or the parallel word), and finally the post-post-critical region (i.e. the two words after *too* or the parallel word).

For the critical region, the main effect of trigger was –0.02 log-ms (89% CrI: (–0.04, –0.007)), which suggests that there was a difference between the *nil* and *too* conditions, in particular, the *nil* condition took longer to read than the *too* condition. The main effect of distance was –0.02 log-ms (89% CrI: (–0.04, –0.008)), which means that the participants read the word faster when the distance increased. The interaction effect was 0.003 log-ms (89% CrI: (–0.01, 0.015)). The mean value of the interaction being very close to 0 and the spread of the CrI across both positive and negative values suggest that its direction is inconclusive.

Using the nested model, we further inspect the effect of distance in the *too* and the *nil* conditions. Within the *too* condition, the effect of distance is –0.02 log-ms (89% CrI: (–0.04, –0.001)), which means that the *long* condition was read faster than the *short* condition. The effect of distance within the *nil* condition was –0.03 log-ms (89% CrI: (–0.05, –0.006)), which also means that the *long* condition was read faster than the *short* condition.

Figure 3 shows the posterior probability distribution of the relevant parameters in the post-critical region. The main effect of trigger was –0.02 log-ms (89% CrI: (–0.04, –0.004)), which again suggests that there was a difference between the *nil* and *too* conditions, and that the *nil* condition took a little longer to read than the *too* condition. The main effect of distance was –0.02 log-ms (89% CrI: (–0.03, –0.004)), which suggests that participants read the word faster when the distance increased. The interaction effect was 0.02 log-ms (89% CrI: (0.005, 0.031)). The positive interval shows that the negative difference in distance, when comparing the *long* and *short* conditions, was diminished and maybe even reversed in the *too* condition.

In the nested model, the effect of distance within the *too* condition was 0.001 log-ms (89% CrI: (–0.02, 0.02)), that is, there was no difference in reading times between the conditions. In the same model, the effect of distance within the *nil* condition was –0.04 log-ms (89% CrI: (–0.06, –0.02)), which means than the *long* condition was read faster than the *short* condition.

Finally, for the post-post-critical, region the main effect of trigger was –0.01 log-ms (89% CrI: (–0.03, –0.002)) which, as in the previous regions, suggests that the *nil* condition took longer to read than the *too* condition. The main effect of distance was 0.005 log-ms (89% CrI: (–0.007, 0.02)), which suggests almost no distance effect. The interaction effect was positive, just as in the post-critical region, even though the credible interval included zero this time (mean 0.008 log-ms, 89% CrI: (–0.003, 0.02)).

In the nested model, we see that the effect of distance within the *too* condition was 0.01 log-ms (89% CrI: (–0.005, 0.03)), which means that the *short* condition was read faster than the *long* condition, even though the null effect is not excluded among possible values in the credible interval. The effect of distance within the *nil* condition was –0.003 log-ms (89% CrI: (–0.018, 0.012)), which means that there was no difference in reading times between the conditions.

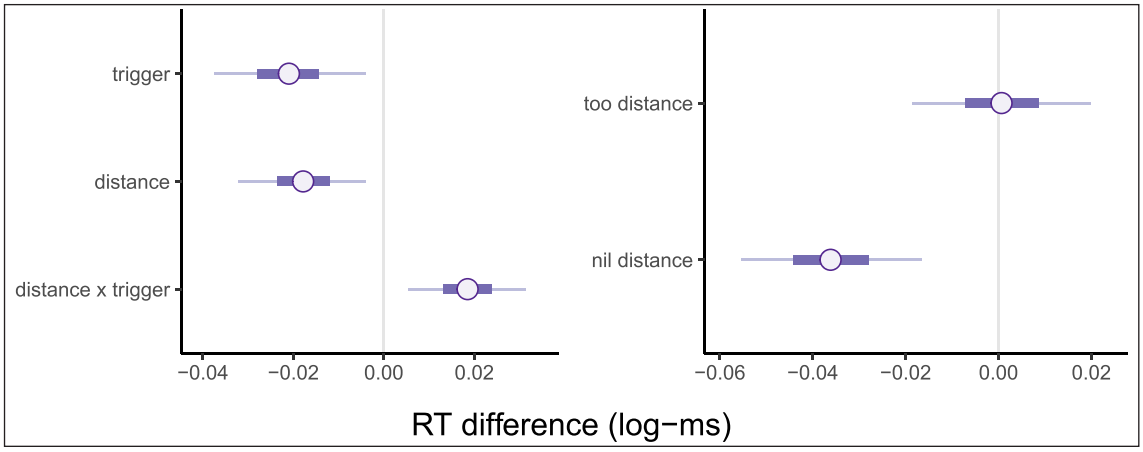

Figure 3: The posterior probability distribution of the parameters of interest at the post-critical region, in Experiment 2. The thicker lines delimit 50% probability intervals, the thinner lines delimit 89% probability intervals. The point signifies the mean value. The plot on the left shows main effects and the interaction, the plot on the right shows effects of distance nested within a given trigger level.

4.7 Discussion

The posterior distribution of the trigger was (in the 89% credible interval) negative in the post-critical region. This suggests that the *nil* condition took longer to read than the *too* condition, possibly because the trigger was more predictable than the corresponding element in the *nil* condition (an adverb). The main effect of distance was also negative in the same region, which suggests that the post-critical region was read faster when the distance was greater. Crucially, the effect of distance interacted with the trigger: there was a positive interaction in the post-critical region. As we mentioned, one way to interpret the positive interaction is that it shows that the negative effect of distance is diminished and maybe even reversed in items with *too*. The pattern of the interaction is easier to understand when we turn to the nested model, shown in **Figure 3**. We see that there was a speed-up in reading due to distance in the *nil* condition, but no such effect in the *too* condition. In other words, the interaction comes about due to the interplay of two effects: the speed-up effect due to distance in the *nil* condition (longer items are read faster in the post-critical region), and the absence of this speed-up in the *too* condition.

We interpret the finding as follows. First, the fact that the *long* condition is read faster than the *short* condition in the *nil* case is in line with the general observation that readers speed up during reading texts or experimental items, as seen in experimental research (Ferreira and Henderson, 1993; see also Vasishth & Lewis, 2006, p. 780) and the analysis of psycholinguistic reading corpora (Demberg & Keller, 2008; Dotlačil, 2021). It is also in line with the fact that we observe a (general) speed-up on the critical region in this experiment. Second, in contrast to the *nil* condition, something nullifies the speed-up in the *too* condition. This, we submit, is the

retrieval process due to *too*. Since readers have to resolve the presupposition, they have to recall the antecedent. Such a recall is harder and slower in the *long* condition compared to the *short* condition. Under this interpretation, our findings would represent a novel instance of semantic inhibitory interference (Parker & Phillips, 2016; Van Dyke, 2007), one that is observed and can be studied in texts and long discourses, not just sentence-internally.

We can think of three possible explanations for the proposed slow-down in *too* compared to the baseline (the *nil* condition). First, it is possible that readers retrieve a distractor and subsequently correct their misretrieval, which takes extra time. Since misretrievals are more likely to happen with increased distance, the discussed reading times pattern is observed. Alternatively, readers might not correct their wrong retrieval but stick to it and consider the presupposition as unresolved. This would result in an infelicitous sentence, which would likely be read slower compared to a fully grammatical, felicitous sentence. Finally, it is possible that with greater distance, the antecedent decays in memory, resulting in longer retrieval time.

It should also be noted that the interaction effect persists to some extent at the post-post-critical region. If this is related to the observed effect, this might suggest that the retrieval difficulty affected the reading process for longer than just immediately after retrieval.

We should mention one possible confound in this study. As we explained in Section 1, additive presuppositions depend on information structure. In particular, the associate of the additive trigger has focus. Given that our stimuli are not spoken, there is no guarantee the information structure we intended for our stimuli matches how participants perceived them. Take, for instance, the following example:

(11) The cook is a dancer and the waiter, who is a boxer, dances too, ….

Here, we assume that when participants read this sentence, they interpret *waiter* as being in focus. The additive then associates with the subject. Potentially, however, participants read (11) with focus on the verb *dances*. In that case, the presupposition becomes that the waiter has properties besides being someone who dances. Given that such properties are provided in the appositive clause intervening between the subject and the verb, our manipulation of distance would stop working under this interpretation. This is even more so because this interpretation is fully unavailable in the short distance condition, where no alternative properties of the subject are given.

While we acknowledge that this alternative understanding of our stimuli is available, we doubt that it played a major role in the results of Experiments 1 and 2. This is because informal intuitions that we gathered about our stimuli suggest that our own intended understanding is much more prominent than the understanding with focus on the verb. However, to be sure, the follow-up experiments were designed to avoid this possible confound.

5. Experiment 3: Acceptability judgment task

5.1 Design

Experiments 3 and 4 were designed to further investigate the retrieval of the antecedents of additives. We wanted to see whether the effect of distance on retrieval can be observed by manipulating the number of sentences in a discourse, not by manipulating the number of words in an appositive, i.e. a non-restrictive relative clause, as was done in Experiments 1 and 2. One reason to consider this is that a non-restrictive relative clause might be processed differently from, and independently of, the rest of the discourse (Dillon et al., 2014), adding unnecessary complications and orthogonal issues to the first two experiments. Another reason is that this allows us to exclude the possible confound in Experiment 1 and 2, that is, the possibility that in the *long* condition, the presupposition was resolved with a different information structure than in the *short* condition.

In the same way as in Experiment 2, we used two factors in Experiment 3: distance and trigger. Two levels of distance were crossed with two levels of trigger, resulting in a 2 × 2 within-participant design.

The distance factor was manipulated by changing the number of clauses between the items. In the *short* case, there was 1 additional clause introduced between the first sentence and the critical sentence. In the *long* case, there were 3 such clauses, see (*short*) and (*long*) in *too* and *nil* below. The sentences were always in the past tense.

too

(*short*) Anne came first in her race. She burst into tears. Caroline won too, setting the lap record.

(*long*) Anne came first in her race. She raised her hands. She screamed out loud. She burst into tears. Caroline won too, setting the lap record.

nil

(*short*) Anne fell down in her race. She burst into tears. Caroline won gold, setting the lap record.

(*long*) Anne fell down in her race. She raised her hands. She screamed out loud. She burst into tears. Caroline won gold, setting the lap record.

The way these discourse fragments were constructed permits us to treat VPs as predicates.[6] Then, the contrast of distance can be interpreted as the number of these VP-predicates that appear in the *short* and *long* conditions. In the case of (*short*), only one extra clause, introducing one such extra predicate, appeared between the predicate with *too* (*won too*) and the antecedent (*(Anne) came first in her race*). In case of (*long*), three extra sentences, introducing three extra predicates, interfered between the predicate with *too* and the antecedent. We hypothesised that the number of predicates between the antecedent and the trigger should influence how quickly the participants will be able to retrieve the information required by the trigger.

[6] Obviously, predicates cannot be equated with VPs in the general case. See the discussion in 2.3.

The trigger factor was manipulated by replacing the trigger *too* with a non-triggering word, often a noun, as in the illustrative item given above. The corresponding word in the *nil* condition (*gold* above) had to be of similar length as *too* (between 3 and 5 characters). As in Experiments 1 and 2, the items in the *nil* condition also differed in the first sentence. Care was taken to ensure that sentences that were changed between the conditions had similar lengths and numbers of words.

Notice that the items above no longer allow for the possible confound we discussed in connection with Experiment 2. There, some stimuli had more than one candidate for the additive's associate. In this experiment, however, the only felicitous associate for *too* is the subject. In particular, assuming the predicate *won* as the associate would make the presupposition unresolved in both *short* and *long* contexts. What's more, the subject of the target sentence is easily recognized as a contrastive topic. In the first two sentences, *Anne* is the topic. The shift of subject in the target sentence means that *Caroline* becomes a contrastive topic (see Büring, 1999, section 1.3). As such, it is the natural associate of the additive (Krifka, 1998; Sæbø, 2004).

We created 24 items of the type described above. The experiment furthermore included 24 filler items. These were simple and short grammatical discourses of a similar form as the target items. Additionally, 10 control items were added to the stimuli. These items were short discourse fragments which violated various semantic and syntactic constraints. They were similar to but not the same as, the ones used in Experiment 1.

5.2 Participants

Participant recruitment and compensation were carried out in the same way as in Experiment 1. 32 participants (female = 14 (44%), male = 18 (56%)) were recruited.

5.3 Procedure

The procedure was very similar to Experiment 1. Some modifications were needed to accommodate the different design of the experiment. The experiment consisted of 4 conditions so the 24 target items were distributed across 4 lists and participants were assigned at random to each of these lists. Each participant saw 6 items from every condition.

5.4 Predictions

The predictions are the same as in Experiment 1, with the caveat that the stimuli are discourse fragments and not sentences.

5.5 Data analysis and results

No participants were removed from the data. In the same way as in Experiment 1 we fit a cumulative ordinal Bayesian model to the data. All the model details were the same as in

Experiment 1. The distance was sum-contrast coded with *long* conditions coded as 1 and *short* conditions coded as –1. The trigger was sum-contrast coded (*nil* = –1, *too* = 1). Fixed effects of interest were: distance, trigger and their interaction.

We used regularising prior distributions for slopes and intercepts. For the intercept, we used a normal distribution with $\mu = 0$ and $\sigma = 2$. For the slopes, we used a normal distribution with $\mu = 0$ and $\sigma = 0.5$. See also Experiment 1 for a more extensive discussion of the prior structures. We report the results on the log-odds scale.

The 89% credible intervals for the slopes of the interest are summarised on **Figure 4**.

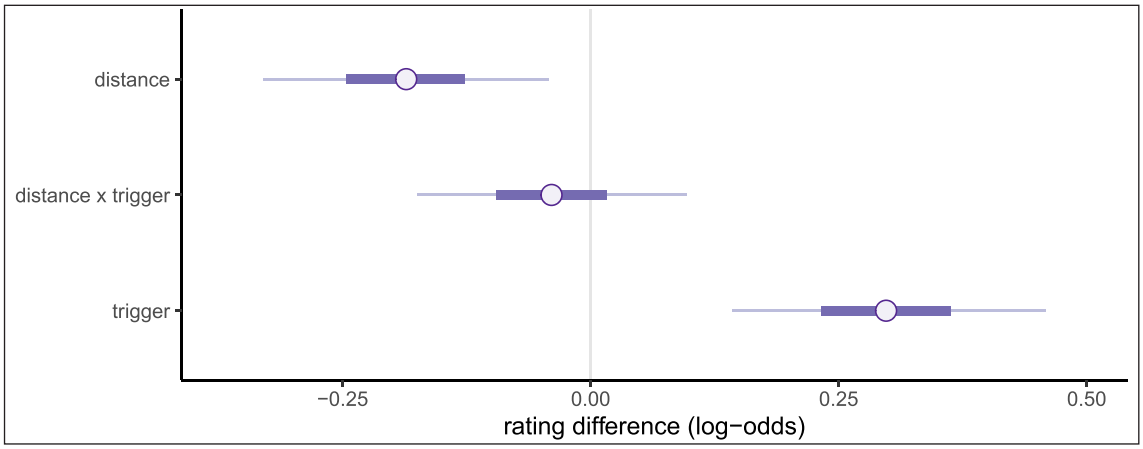

Figure 4: The posterior probability distribution of the parameters of interest in Experiment 3. The thicker lines delimit 50% probability intervals, the thinner lines delimit 89% probability intervals. The point signifies the mean value.

As in the previous experiments, the crucial parameter here is the slope of the interaction. The posterior distribution of the slope of the interaction (89% CrI = (–0.2, 0.1), mean = –0.04, sd = 0.09) suggests that the effect of distance on acceptability does not clearly differ between the trigger conditions.

The estimate of the distance parameter indicates that the longer items were judged as less natural (89% CrI = (–0.3, –0.04), mean = –0.2, sd = 0.09).

In the case of the trigger, the fragments with *too* were more natural to the participants than those without it (89% CrI = (0.1, 0.5), mean = 0.3, sd = 0.1).

6. Experiment 4: Self-paced reading

6.1 Design

We conducted a self-paced reading study using the same design as in Experiment 3. Some nonessential changes were applied to the stimuli, and the control items were removed. Out of 48 items, 28 (17 target items and 11 filler items) were followed by a comprehension question.

The list of items is provided at the OSF repository.

6.2 Participants

79 participants (female = 45 (57%), male = 31 (39%), other/don't want to say = 3 (4%)) were recruited for the self-paced reading study using Prolific (https://www.prolific.co).

The participants were compensated 3.00 GBP for taking part in the experiment. The experiment took, on average, 16 minutes and 46 seconds to finish; therefore the participants would get 10.7 GPB per hour, on average. The participants were self-reported English native speakers.

6.3 Procedure

The recruitment of the participants, randomization, and data collection mimicked that in Experiment 2. The experiment was conducted in a noncumulative self-paced reading moving window paradigm, as in Experiment 2. Since the stimuli were constructed of quite long text fragments, we inserted line breaks manually to ensure that the regions of interest would not end up before or across the line break.

6.4 Predictions

The predictions are the same as in Experiment 2, with the caveat that the stimuli are discourse fragments and not sentences.

6.5 Data analysis

Every participant answered more than 70% of comprehension questions correctly. For this reason, no participant was removed. Time measurements which were under 50 ms or above 3,000 ms were removed from the results. Similar data trimming was done, for example, by Futrell et al. (2021). This resulted in the removal of 0.46% of the data.

Data analysis of the final data from 79 participants was similar to the data analysis conducted for Experiment 2.

We fitted a Bayesian mixed-effects model to the response time data, measured at the critical, post-critical, and post-post-critical regions. The experimental factor *trigger* was coded using sum contrasts, with the *nil* condition as –1 and the *too* condition as 1. The factor *distance* was also coded using sum contrasts, with the *short* condition coded as –1 and *long* condition as 1. Both of these variables were used as fixed effects in the model. Participants and items were entered as random effects.

We also fit two additional models to the data. The first of these was the same nested model as in Experiment 2. With the second of those, we wanted to check for the effect of distance in

regions other than the critical and post-critical. Here, we first pooled the data from the regions starting at the critical region –5, i.e. the first region at which the distance manipulation could be detected, up to the pre-critical region and the post-post-critical region (excluding the critical and post-critical region). The post-post-critical was the last region at which all the items had measurements. To these data, we fit a similar model as before, but without the trigger predictor. After this, we also examined the effect of distance at each of those particular regions. This was done to check for secondary support for our initial explanation, i.e. that the retrieval of the antecedent is erasing the general speed-up. The reasoning was that, if there is a general speed-up, we should also be able to observe it in other regions besides the critical and the post-critical regions. These models used 4 sampling chains with 6,000 samples drawn from each chain. 2,000 samples from each chain were discarded for warm-up; hence, each model had 16,000 samples available for the analysis.

The priors used were the same as the ones used Experiment 2. For the intercept, we used a normal distribution with $\mu = 6$ and $\sigma = 1.5$. For the slopes, we used a normal distribution with $\mu = 0$ and $\sigma = 0.1$. The prior distribution used for the standard deviations of the random effects and for the residual standard deviation was a truncated normal distribution with $\mu = 0$ and $\sigma = 1$. For the random effects correlation between the intercept and the slope we used the LKJ distribution (Lewandowski et al., 2009; Stan Development Team, 2021) with $\eta = 2$. See also Section 4.5 for a more extensive discussion of the prior structures.

Model estimates are reported on the log-ms scale.

6.6 Results

The descriptive summaries of the reading times across conditions for one word before and two words after the critical region (the additive *too* and the corresponding word in the *nil* condition) are shown on **Figure 5**.

At the critical region, the main effect of trigger was –0.02 log-ms (89% CrI: (–0.04, –0.008)). This suggests a negative difference between the conditions, i.e. the *too* condition was read faster than the *nil* condition. The main effect of distance was –0.009 log-ms (89% CrI: (–0.02, 0.003)). Since the probability density of the 89% credible interval is on the both sides of zero, one cannot conclusively infer the direction of the effect from the estimate. However, the region has much more density below zero which means that it is more likely that the participants were reading the word faster in the *long* condition than in the *short* one. The interaction effect was 0.005 log-ms (89% CrI: (–0.01, 0.02)). This suggests that the difference between the *long* and *short* conditions was smaller in the *too* conditions than in the *nil* conditions, although a caveat similar to the distance results applies here also. All the results observed at the critical region had the same direction as in Experiment 2. The magnitude of the results was also similar excluding the effect of distance, which, in Experiment 2, was more negative.

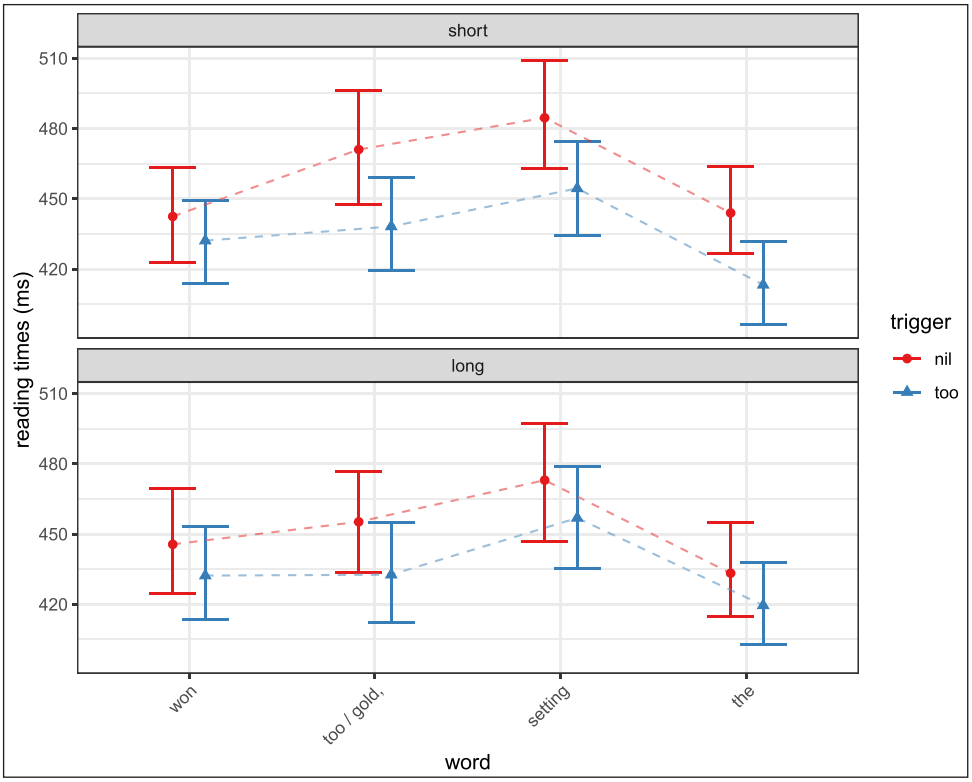

Figure 5: Descriptive summary of reading times in Experiment 4. The points connected by dashed lines denote mean reading times across conditions. These were computed by taking an average of all the data points for each condition, i.e. without grouping by participants or items. The bars represent bootstrapped 0.95 confidence intervals.

Let us turn to the nested model for the critical region. The model shows that the effect of distance within the *too* condition was –0.006 log-ms (89% CrI: (–0.02, 0.01)) and that the effect of distance within the *nil* condition was –0.01 log-ms (89% CrI: (–0.03, 0.007)). The results suggest that *long* was read faster than *short* in both conditions, even though the credible intervals include zero and, thus, the speed-up should not be taken as conclusive. Both of these effects were much closer to 0 than in the Experiment 2, but the fact that more probability density is below 0 aligns with the results observed in Experiment 2.

At the post-critical region, the main effect of trigger was –0.02 log-ms (89% CrI: (–0.03, –0.005)), i.e. the *too* condition was read faster than the *nil* condition. The main effect of distance was –0.01 log-ms (89% CrI: (–0.02, 0.001)). As in the case of the critical region, the probability density of the 89% credible interval is on both sides of zero, but most of the density is below zero. Therefore, it is more likely that the participants were reading the word faster in the *long* condition than in the *short* one. The interaction effect was 0.01 log-ms (89% CrI: (0, 0.02)). This suggests that the difference between the *long* and *short* conditions, which was negative as the main effect, was diminished or even (as we indeed will see) reversed in the *too* condition.

The effects of trigger and interaction were virtually the same as in Experiment 2, the effect of distance was diminished compared to Experiment 2. The posterior probability distribution of the parameters is summarized in **Figure 6**.

The interaction can be more easily understood in the nested model, which shows that the effect of distance within the *too* condition was 0.01 log-ms (89% CrI: (–0.007, 0.03)). That is, it is more likely that the *too* condition took more time to read in the *long* than in the *short* condition. However, since the credible interval crosses zero this interpretation should not be taken as conclusive. The effect of distance within the *nil* condition was –0.02 log-ms (89% CrI: (–0.04, –0.006)), which means that the *long* condition was read faster than the *short* condition. Both of these effects were more positive than in Experiment 2.

For the post-post-critical region, the results are very similar. The main effect of trigger was –0.02 log-ms (89% CrI: (–0.04, –0.008)), that is, the *too* condition was read faster than the *nil* condition. The main effect of distance was –0.004 log-ms (89% CrI: (–0.01, 0.006)). The interaction effect was 0.007 log-ms (89% CrI: (–0.003, 0.02)). The effects of trigger and interaction were virtually the same as in Experiment 2, the effect of distance was more negative compared to Experiment 2.

In the nested model of the post-post-critical region, the effect of distance within the *too* condition was 0.001 log-ms (89% CrI: (–0.01, 0.02)), that is, there was no clear difference in reading times between the distance conditions. The effect of distance within the *nil* condition was –0.01 log-ms (89% CrI: (–0.03, 0.003)), which means that it is more likely that the participants were reading the *long* condition faster than the *short* condition, but since the credible interval crosses zero, the interpretation is inconclusive. These effects were the only ones that were reversed compared to the results from Experiment 2. The effect of distance within *too* condition was much smaller in this experiment, and the effect of distance within *nil* condition was much larger in this experiment.

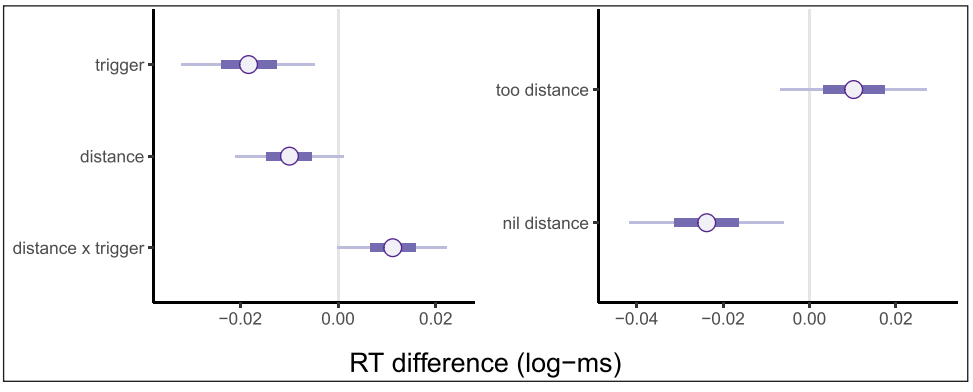

Figure 6: The posterior probability distribution of the parameters of interest, from the model fit to the post-critical region in Experiment 4. The thicker lines delimit 50% probability intervals, the thinner lines delimit 89% probability intervals. The point signifies the mean value.

6.7 Discussion

The 89% credible interval of the estimate of the interaction parameter is predominantly situated on the positive side in the critical and the post-post-critical regions, and it is fully situated on the positive side in the post-critical region. The finding in the post-critical region replicates what we observed in Experiment 2. In the nested comparison of the post-critical region, we see that the interaction is driven by the speed-up in long items in the *nil* condition, which is crossed with a slowdown in long items in the *too* condition. The slowdown due to distance in the *too* condition, however, was not fully conclusive, since the credible interval crossed zero. In Experiment 2, we suggested that this can be explained as an interplay of two effects: readers speed up during reading, hence the speed-up due to distance in the *nil* condition, whereas the retrieval process is more effortful when the antecedent being retrieved is further away, hence the lack of speed-up, or even slowdown, in the *too* condition. We assume the same explanation accounts for the interaction in this experiment.

We can explore this explanation further since the *short* and *long* conditions were matched in several words preceding the critical region. Somewhat abstractly, we can represent the situation in this experiment as in (12). Sentence1, Sentence2 and Sentence3 are matched across the *short* and *long* conditions. Sentence3 includes the crucial manipulation, the additive particle *too*, which triggers retrieval. We can go up to 5 words back to the beginning of Sentence2 to explore the effect of distance on matched words before *too* was read.

(12) (*short*) Sentence1. Sentence2. Sentence3.
(*long*) Sentence1. …**extra material** …Sentence2. Sentence3.

If our explanation is right, we expect to find a speed-up in these regions. In fact, the findings are mixed. When we pool regions in Sentence2 and Sentence3 before the critical region, but including the post-post-critical region, we see a negative effect, corresponding to the expected speed-up: –0.002 log-ms (89% CrI: (–0.01, 0.008)). However, as we can see, the credible interval does not exclude zero. When we look at individual words, we see a clear negative effect (i.e. faster reading in longer distance) at region –4, four words back from the critical region: –0.02 log-ms (89% CrI: (–0.04, –0.006)). It is probably relevant that in this region a clause (Sentence2) ended in 18 out of 24 items, and there was a line break in the majority of test items after this word. In this respect, it might be worth pointing out that the critical region, which is the next region to show a quite clear speed-up effect of distance, is also the end of a clause, which is signalled by a comma, and followed by a gerund clause. However, it is not clear why we would see a spillover of the speed-up effect to the post-critical and the post-post-critical regions, but no comparable spillover from the region –4. Thus, while we have some evidence that the speed-up effect of distance can be found outside of the critical and the post-critical regions, the evidence is not entirely conclusive, since only one region shows this effect clearly.

The results of the post-critical region replicate the results from Experiment 2 even though Experiment 4 looks at retrieval across a series of sentences, while Experiment 2 investigated the effect of distance within a clause. This suggests that the memory processes involved in the retrieval of the antecedent of *too* might be similar both at the sentence level and at the discourse level.

As can be seen in **Figure 5**, the data are noisier than in Experiment 2. This can be related to the stimuli being discourse fragments instead of simple sentences.

7. A pooled model

Since the design of all the experiments was similar, we decided to take advantage of the Bayesian paradigm and fit a model to the data pooled from all experiments and grouped by task type. We wanted to check what trend the data point to when taken together. Furthermore, we calculated Bayes factors on the pooled data to compare two hypotheses of interest. There are some important caveats to be made here. The Bayes factor is somewhat controversial (see Gelman et al., 2003, pp. 182–184; Navarro, 2019). One reason for this status is that the Bayes factor is "highly sensitive to, and crucially depends on, prior assumptions about model parameters" (Schad et al., 2021, p. 6). We tried our best to be cautious, but the reader should keep these possible shortcomings in mind. We think that the Bayes factor estimation is useful, but it should be mostly treated as an addition to the main content of the analysis. For a good overview of the Bayes factor methodology, see Schad et al. (2021).

In the model estimation used for the calculation of the Bayes factor for the self-paced reading experiments, we used 4 chains, with 30,000 iterations per chain and 2,000 warmup iterations. The stability of the estimates obtained was checked by fitting the models (the null model and the alternative model) 4 times, estimating the marginal likelihood for each of these fits via bridge sampling, and, finally, computing the Bayes factor from these estimates. A similar approach was used in the case of the acceptability judgment experiments, but these models used 4 chains with 15,000 iterations per chain and 2,000 warmup iterations. The stability of these estimates was checked by completing the estimation procedure 5 times.[7]

In the first hypothesis, which we label H1, it is assumed that next to the effect of the trigger (*too* vs. *nil*) and the effect of distance, the interaction term *trigger* × *distance* as a fixed effect influences reading times. In the second hypothesis, which we label H0, it is assumed that the interaction term does not influence reading times. Using the common notation of the R model formula, we write these two hypotheses as the following two formulae:

[7] There were also other differences. In particular, for the SPR experiments stability of the estimates was checked for each value of σ separately, while for the AJ experiments this was done only once. See the OSF repository for details.

(13) [H1] log(rt) ~ 1 + distance * trigger +
(1 + distance * trigger | participant) +
(1 + distance * trigger | item)

[H0] log(rt) ~ 1 + distance + trigger +
(1 + distance * trigger | participant) +
(1 + distance * trigger | item)

7.1 Data preparation

For each experiment type (i.e. acceptability judgment and self-paced reading), we took all the collected data and treated them as if they were gathered in one experiment in which subsets of participants saw subsets of items.

For the self-paced reading tasks, we only considered data from the post-critical regions. To such datasets we fit a model with random intercepts and slopes for participants and items. The fixed effects were *distance*, *trigger* and their interaction. The predictor *trigger* was sum-coded with *nil* = –1, and *too* = 1. The predictor *distance* was the number of words between the antecedent and the trigger z-normalized, i.e. rescaled to a mean of 0 and standard deviation of 1.

For the acceptability judgment experiments, we treated the data in a similar way, but since the distance factors differed between the experiments (4 levels in the first experiment and 2 levels in the second), we took a subset of items from the first experiment (levels labelled 0 as *short* and 8 as *long*) and fit models to this subset. The distance was sum-coded with the *long* distance coded as 1 and the *short* distance as –1.[8]

7.2 Results

First, we consider the self-paced reading data. We report the estimates of the models with the interaction. The main effect of the trigger was –0.02 log-ms (89% CrI: (–0.03, –0.008)). The negative credible interval shows that the post-critical regions are more likely read faster in the *too* condition than in the *nil* condition. The main effect of distance was –0.016 log-ms (89% CrI: (–0.03, –0.006)). This means that the participants read the post-critical region faster when the distance was greater. Finally, the interaction effect was 0.017 log-ms (89% CrI: (0.008, 0.03)). This shows that the negative effect of distance was diminished, or even reversed, in the *too* condition. The interaction estimate replicates the findings from individual experiments and supports our hypothesis that increased distance is problematic for presupposition resolution, because the retrieval of the antecedent is more effortful.

[8] A more complete model for the pooled data would also account for the source of the data, i.e. which experiment the data comes from. Such a model would assume an additional group *experiment* and would estimate the parameters for that group also. However, when we tried this, we encountered issues with the convergence of the model, and hence we decided to use the model described in the text instead.

For the acceptability judgment data, the main effect of the trigger was 0.2 log-odds (89% CrI: (0.08, 0.35)), suggesting that the *too* condition was judged as more natural than the *nil* condition. The main effect of *distance* was –0.3 log-odds (89% CrI: (–0.5, –0.2)). This means that the participants judged longer sentences as worse on average than short sentences. The interaction effect was –0.08 log-odds, showing that the effect of distance was greater (more negative) in the *too* condition, but the credible interval included zero (89% CrI: (–0.19, 0.019)).

We now turn to the comparison of the two hypotheses H0 and H1 using the Bayes factor. Since, as mentioned above, the Bayes factor is sensitive to assumptions about prior distributions (Gelman et al., 2003; Navarro, 2019; Schad et al., 2021), we provide a sensitivity analysis. This means that we check several priors for the part of the model in which H0 and H1 differ, i.e. the interaction term distance $\times$ trigger.

The summary of the sensitivity analysis for Bayes factor BF_{10} regarding the self-paced reading studies is provided in **Table 1**. In this sensitivity analysis, the prior distribution for the interaction term is normal, with μ centered at zero and σ one value from the list 0.01, 0.05, 0.10, 0.20, 0.50. We chose these values for σ since small values (0.01 and 0.05) restrict the interaction term within the most plausible range and should thus lead to a BF_{10} that shows evidence in favor of the model with the interaction term, H1 (provided the data support H1). The remaining values (0.1, 0.2, 0.5) are unreasonably large and overestimate the effect size of the interaction term, and, thus, it is likely that somewhere among these values, we will observe that BF_{10} should become smaller than 1 and thus reveal evidence in favor of H0, the model without the interaction term.

A similar reasoning could be applied to the sensitivity analysis of the Bayes factor regarding the acceptability judgment task. Again, we decided to use a normal prior distribution, with μ set to 0 and σ taking values from the list 0.005, 0.009, 0.05, 0.1, 0.3, 0.5. We assumed that the smaller values of σ should provide more evidence for the model with the interaction term and the larger values, for the model without the interaction term, but note that here we do not really have a good reference point for what effect size should be expected.

In **Table 1**, we see that when the standard deviation of the prior distribution is ≤ 0.2, Bayes factor estimates show evidence in favour of the model with the interaction, and only when the standard deviation is quite large, $\sigma = 0.5$, is BF_{10} smaller than 1, showing evidence in favour of the null model (the model without the interaction term). This decrease of evidence in favor of H1 is not surprising, given that other studies of interference effects in memory retrieval found clear evidence in favor of H1 only when the effect was small (Schad et al., 2021).

Among the values considered, there is a maximum, ($BF_{10} = 13.3$), when the standard deviation is 0.01, i.e. the observed data in the pooled model are 13.3 times more likely to occur under H1 than H0 in this case. This size of BF_{10} is sometimes labeled as moderate-to-strong evidence in favor of the full model (van Doorn, et al., 2021).

Table 1: Bayes factors (BF_{10}) for the self-paced reading studies for the models in which the prior distribution of the interaction term was normal distribution, with the mean centered at zero and the standard deviation a value from the list 0.01, 0.05, 0.10, 0.20, 0.50.

σ	BF_{10} (mean)	BF_{10} (individual runs)
0.01	13.3	10.9, 12.1, 14.8, 15.4
0.05	7.8	5.8, 8.5, 8.5, 8.7
0.10	4.7	4.1, 4.6, 4.6, 5.5
0.20	2.1	1.87, 1.85, 2.26, 2.32
0.50	0.82	0.76, 0.8, 0.8, 0.91

Table 2: Bayes factors (BF_{10}) for the acceptability judgment studies. Listed are the models in which the prior distribution for the interaction term was normal distribution, with the mean centered at zero and the standard deviation a value from the list 0.005, 0.009, 0.05, 0.1, 0.3, 0.5.

σ	BF_{10}
0.005	1.129
0.009	1.045
0.050	1.041
0.100	0.985
0.300	0.468
0.500	0.347

Regarding the acceptability study, we see in **Table 2** that no values of σ provide evidence for the model with the interaction. In fact, the estimates for most values of σ are close to 1, which means that there is no change in evidence for either of the models, as far as acceptability judgment studies are concerned.

7.3 Discussion

In the pooled reading studies, we see that the 89% credible interval for the interaction parameter lies on the positive side of 0, indicating that for the pooled model, it is likely that the effect is positive. Moreover, the Bayes factor estimates indicate that there is clear evidence (moderate-to-

strong in the range of most reasonable priors) in favor of the H1 model. There was no evidence in favor of the interaction model in case of the pooled acceptability studies, that is, the detectable slowdown in reading does not translate to decreased acceptability in our experiments.

The sensitivity analysis of the Bayes factor shows the relationship between the standard deviation of the prior distribution on the interaction parameter, and the strength of the evidence for the full model (the model with the interaction term).

The interaction term expresses the difference between the effect of distance in the *too* condition and the effect of distance in the *nil* condition. The model fitted to the pooled data predicted the mean estimate of the interaction term as 0.0173 log-ms. It is easier to interpret these estimates on the ms scale. At the intercept estimate of the model, this corresponds to a difference of 14 ms. The difference is calculated as follows:

$$(14) \quad \exp(\textsc{Intercept} + \beta_{int}) - \exp(\textsc{Intercept} - \beta_{int}) = \exp(6 + 0.0173) - \exp(6 - 0.0173)$$

This means that at the intercept (here, the grand mean reading time), we expect 14 ms of difference per one unit of distance. Recall that the distance is measured in the scaled number of words between the beginning of the text and the trigger. On this scale, one unit of distance corresponds to 6.2 words.

8. General discussion

8.1 The size and source of the effect

The results of these experiments provide evidence that reading is slowed down when the antecedent is further away from the trigger, the additive particle *too*. However, given that retrieval of the antecedent is needed, and given that it seems clear that this retrieval should be more taxing with increased distance, it should not be difficult to find evidence for this distance effect. Yet, the effects in both experiments are rather small. The strength of this evidence, as quantified by the Bayes factor, is conditional on the prior distribution for the slope of the interaction used in the model. We saw moderate to strong evidence for models with prior distributions that assume rather small effects sizes.

Notice, however, that memory retrieval effects seem to be small in general (Jäger et al., 2017). Their small effect size is also what makes them difficult to establish. As an example, consider facilitatory interference effects in antecedent-reflexive constructions. A study by Dillon et al. (2013) found no facilitation, but a later experiment which collected data from many more participants (Jäger et al., 2020) did find some such effect.

The size of the effect also provides clues as to how we should reason about its source. It is usually assumed that retrieval can happen either via a serial search mechanism or via a direct-access, cue-based mechanism (McElree et al., 2003). The former is a costly mechanism; the parser has to attend to every possible item in search for a match. To believe that a similar

mechanism is responsible for retrieval in the experiments presented, we would expect a drastic slowdown with every additional candidate. We did not observe such a slowdown. Rather, the approximate size of the effect is comparable to the findings on retrieval in syntax. Assuming that only these two possibilities are on the table, we are compelled to conclude that the mechanism responsible is direct access.

The slowdown observed in the experiments could be interpreted as an effect of decay: the greater distance resulted in a slower retrieval of the relevant information, because its representation in memory weakened with time. Interpreting the findings in terms of decay is compatible with cue-based theories of retrieval, which can be cast in the framework in which temporal decay is one of the factors affecting retrieval time (e.g. ACT-R Anderson & Lebiere, 1998; Anderson, 2007).

Another possibility is that the effect is driven by a different kind of cue from the ones we assumed so far. The experiments assumed that all the cue-value pairs other than the ones directly manipulated remained constant across the conditions. However, there are cue-value pairs the status of which might change with distance. For example, the salience of the licensing element might decrease with distance.[9] If the retrieval of the antecedent is guided by a cue {*salient*}, and salience diminishes with distance, we would expect the parser to have trouble retrieving the target.

8.2 Additive presupposition between retrieval and expectation

We started this work arguing that additive presuppositions trigger a retrieval process, on the basis of the fact that language users have the capacity to judge the felicity of additive *too*, given the presence or absence of appropriate material earlier in the discourse. That is, the contrast between (15) and (16) shows that the processing *too* relies on recall of earlier parts of the discourse.

(15) Ann lives in Paris. Sue lives in France, #(too).

(16) Ann lives in Brussels. Sue lives in France, (#too).

The way we have presented the reliance on memory so far is in line with the treatment of presupposition as an anaphora-like phenomenon, as per quite a few proposals from semantic theory (Kripke, 2009; van der Sandt & Geurts, 1991; van der Sandt, 1992;). However, one aspect of the contrast between (15) and (16) may be relevant for processing, yet does not follow from the anaphoric character of additive presuppositions. Not only is the additive felicitous in

[9] We are grateful to an anonymous reviewer for suggesting this explanation.

(15), it is also obligatory.[10] That is, upon having processed *Sue lives in France* in (15), anything other than an additive will be deemed infelicitous. This obligatory presence of *too* is usually explained along the following lines. If *Sue* is a contrastive topic in the second sentence and the additive is left out, then the presence of alternatives to Sue would trigger the implicature that (out of the contextually salient set of individuals) only Sue lives in France. But this implicature is contradicted by what we just learned about where Anne lives. By presupposing that some alternative is true in addition to the asserted one, exhaustification is cancelled. This is because only one operation can associate with the focus/contrastive topic at a time. So, the presence of *too* precludes exhaustification from taking place and this is why the additive is obligatory in (15) (e.g. Bade, 2016; Krifka, 1998; Sæbø, 2004).

We do not believe this aspect of additive presuppositions drastically changes the retrieval picture, though it may suggest a slightly different narrative about what is behind our results.[11] As explained above, the obligatory nature of *too* makes it very predictable. This predictability is probably supported by various factors, one of which is the parallelism in the items of our first two experiments, as, for example, in (17).

(17) The cook is a dancer and the waiter, who is a great boxer from southern Amsterdam, dances too, I have been told recently.

When processing such a sentence, after arriving at the site of the verb in the second clause (*dances*), participants can retrieve the predicate (*is a dancer*) from the first clause and, as a result, predict *too*. This prediction is likely to result in an increase in reading speed, but such a speed-up is preceded by a retrieval effort, which is less costly for shorter sentences (since retrieval of the previous predicate is less inhibited by the distracting predicates). Because the items in Experiments 3 and 4 do not include the kind of parallelism found in (17), it could be that the prediction of *too* is less profound in those experiments, which could account for the smaller effect.

On this view, one may expect to see the effect of distance earlier, since the retrieval happens at the verb, but it is likely such differences cannot be captured in the self-paced reading paradigm. In short, then, we think that although the predictability of *too* allows for a narrative that is slightly different from the anaphoric processing framing we used elsewhere, the overall predictions are not so different, since the kind of retrieval involved is the same.

---

[10] Note, that there are some exceptions to the obligatoriness of *too*. For instance, *too* is not obligatory in lists: *Sue lives in Paris, Anne lives in Paris, Bo lives in Paris, all my friends live in Paris! Too* can also be skipped when the fragment continues, e.g. … *Sue lives in France but commutes to Rome.* is fine. None of the examples that we looked at in this work are like this, however.

[11] We would like to thank an anonymous reviewer for discussion.

8.3 Slowing down or just not speeding-up?

Our starting point was to combine insights from semantic approaches to additive presuppositions with psycholinguistic findings on memory retrieval. The combination lead us to the hypothesis that an increase in distance between the additive *too* and the antecedent should result in more costly retrieval, which should translate into a slowdown in reading. In two self-paced reading experiments, this slowdown in *too* was observed, but, crucially, the slowdown was observed *with respect to the baseline*, that is, the sentences that lacked the additive. Furthermore, as we saw, an increase in distance caused a speed-up in the baseline. A slowdown in sentences with *too* relative to this baseline means, in the absolute measures studied in the nested comparisons, no speed-up/slowdown (Experiment 2) or a slowdown which, however, is not fully conclusive (Experiment 4).

We suggested that the speed-up in the baseline is expected independently, based on previous findings that show that readers speed up as they proceed in reading text (Demberg and Keller, 2008; Dotlačil, 2021; Ferreira and Henderson, 1993; see also Vasishth & Lewis, 2006, p. 780). Furthermore, we claimed that the fact that we observe a slowdown compared to the baseline but no absolute slowdown, in sentences with *too* should be understood to arise through the combination of two factors. On one hand there is a slowdown caused by more costly retrieval when the distance from the antecedent increases. On the other hand, readers read longer texts and, therefore, they should speed up. The net result is potentially no observable absolute slowdown in sentences with *too*.

Such a trade-off between speed-up due to increased sentence length and slowdown due to harder retrieval was observed in other studies on retrieval and was discussed, among others, by Levy & Keller (2013); Futrell et al. (2020). In those studies, however, the focus was on how specific readers' expectations, modeled in Surprisal Theory, interact with the distance effects observed in retrieval.

One way to strengthen the results of the current study would be to design an experiment in which the distance in *too* does not cause just a relative speed-up but in which a slowdown can be seen irrespective of any baseline. Nicenboim et al. (2016) present an extended discussion of problems that arise when researching distance effects which could potentially be useful in constructing novel experiments that would further strengthen the results in this way.

9. Conclusion

We presented four experiments that investigated the role of the retrieval needed to correctly interpret utterances with the presupposition trigger *too*. The trigger used in the experiments is an example of a linguistic element that forms a semantic dependency, which is not mediated by particular syntactic constructions.

Across the experiments, we found that it takes more time to read sentences with the presupposition trigger *too* when the distance between the trigger and its antecedent is greater, compared to the baseline, which lacked the presupposition trigger. This is compatible with the assumption that *too* triggers retrieval and that retrieval is cue-based, or that the activation of memory traces is subject to temporal decay.

The fact that retrieval is needed to resolve both syntactic and semantic dependencies, and that the retrieval mechanisms seem to show similar properties, suggests that a general-purpose mechanism for memory structuring and memory retrieval could exist in language processing.

Data accessibility statement

The data, materials, and code are available at an Open Science Foundation repository, at: https://osf.io/2xcgu/, DOI: 10.17605/OSF.IO/2XCGU

Ethics and consent

The studies in this article have been conducted after approval by the Faculty Ethics Assessment Committee Humanities (FEtC-H) of Utrecht University under reference number 4058356-01-2019.

Acknowledgements

We thank Sol Lago and two anonymous reviewers for their valuable feedback on this article. The research in this article was supported by an NWO Vrije competitie grant under grant number VC.GW17.112.

The third author was funded by the European Union (the European Research Council grant 101088098, MEMLANG). The views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

Conceptualization: Rick Nouwen, Jakub Dotlačil, Jan Winkowski; Data curation: Jan Winkowski; Formal analysis: Jan Winkowski, Jakub Dotlačil; Funding acquisition: Rick Nouwen; Investigation: Jan Winkowski; Methodology: Jakub Dotlačil, Jan Winkowski; Project administration: Jan Winkowski; Supervision: Rick Nouwen, Jakub Dotlačil; Visualization: Jan Winkowski; Writing — original draft: Jan Winkowski, Jakub Dotlačil, Rick Nouwen; Writing — review & editing: Jan Winkowski, Jakub Dotlačil, Rick Nouwen.

ORCiD IDs

Jan Winkowski 0000-0002-0301-8471

Rick Nouwen 0000-0001-9571-4644

Jakub Dotlačil 0000-0002-5337-8432

References

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

Anderson, J. R. (2007). Cognitive architecture. *How Can the Human Mind Occur in the Physical Universe? 344*. DOI: https://doi.org/10.1093/acprof:oso/9780195324259.003.0001

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Bade, N. (2014). Obligatory implicatures and the presupposition of "too". *Proceedings of Sinn und Bedeutung, 18*, 42–59.

Bade, N. (2016). *Obligatory presupposition triggers in discourse: Empirical investigations of the theories maximize presupposition and obligatory implicatures* [Doctoral Thesis]. University of Tuebingen.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1178. DOI: https://doi.org/10.1037/a0024194

Brasoveanu, A., & Dotlačil, J. (2020). *Computational cognitive modeling and linguistic theory*. Springer (Open Access). DOI: https://doi.org/10.1007/978-3-030-31846-8

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539. DOI: https://doi.org/10.1037/0033-295X.114.3.539

Büring, D. (1999). Topic: Linguistic, cognitive, and computational perspectives. In P. Bosch & R. van der Sandt (Eds.), *Focus* (pp. 142–165). Cambridge University Press.

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

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

Chemla, E., & Bott, L. (2013). Processing presuppositions: Dynamic semantics vs pragmatic enrichment. *Language and Cognitive Processes*, *28*(3), 241–260. DOI: https://doi.org/10.1080/01690965.2011.615221

Chen, S. Y., & Husband, E. M. (2018). Comprehending anaphoric presuppositions involves memory retrieval too. *Proceedings of the Linguistic Society of America*, *3*(1), 1–44. DOI: https://doi.org/10.3765/plsa.v3i1.4288

Chomsky, N. (1957). *Syntactic structures*. De Gruyter Mouton. DOI: https://doi.org/10.1515/9783112316009

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

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*(1), 172–201. DOI: https://doi.org/10.1111/cogs.12227

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. DOI: https://doi.org/10.1016/j.cognition.2008.07.008

Dillon, B., Clifton, C., & Frazier, L. (2014). Pushed aside: Parentheticals, memory and processing. *Language, Cognition and Neuroscience*, *29*(4), 483–498. DOI: https://doi.org/10.1080/01690965.2013.866684

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. DOI: https://doi.org/10.1016/j.jml.2013.04.003

Dotlačil, J. (2021). Parsing as a cue-based retrieval model. *Cognitive Science*, *45*(8). DOI: https://doi.org/10.1111/cogs.13020

Ferreira, F., & Henderson, J. (1993). Reading processes during syntactic analysis and reanalysis. *Canadian Journal of Experimental Psychology*, *47*(2), 247–275. DOI: https://doi.org/10.1037/h0078819

Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, *19*(1), 87–107. DOI: https://doi.org/10.1007/s11050-010-9065-3

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

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The natural stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77. DOI: https://doi.org/10.1007/s10579-020-09503-7

Gelman, A., Carlin, J. B., Stern, H. S., Vehtari, A., Dunson, D. B., & Rubin, D. B. (2003). *Bayesian data analysis*. Chapman; Hall/CRC. DOI: https://doi.org/10.1201/9780429258480

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511790942

Geurts, B. (1999). *Presuppositions and pronouns*. Elsevier.

Geurts, B., & van der Sandt, R. (2004). Interpreting focus. *Theoretical Linguistics*. DOI: https://doi.org/10.1515/thli.2004.005

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, *2000*, 95–126. DOI: https://doi.org/10.7551/mitpress/3654.003.0008

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290. DOI: https://doi.org/10.1207/s15516709cog0000_7

Heim, I. (1982). *The semantics of definite and indefinite noun phrases* [Doctoral dissertation, UMass Amherst].

Huang, Y. T., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, *26*(8), 1161–1172. DOI: https://doi.org/10.1080/01690965.2010.508641

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. DOI: https://doi.org/10.1016/j.jml.2017.01.004

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*, 104063. DOI: https://doi.org/10.1016/j.jml.2019.104063

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. DOI: https://doi.org/10.1037/0096-3445.111.2.228

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, & M. Stokhof (Eds.), *Formal methods in the study of language* (pp. 277–322). Mathematical Centre Tracts.

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Kluwer Academic Publishers.

Kiss, K. É. (1998). Identificational focus versus information focus. *Language*, *74*(2), 245–273. DOI: https://doi.org/10.1353/lan.1998.0211

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

Krifka, M. (1998). Additive particles under stress. *Proceedings of Semantics and Linguistic Theory*, *8*(8), 111–128. DOI: https://doi.org/10.3765/salt.v8i0.2799

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3–4), 243–276. DOI: https://doi.org/10.1556/ALing.55.2008.3-4.2

Kripke, S. A. (2009). Presupposition and anaphora: Remarks on the formulation of the projection problem. *Linguistic Inquiry*, *40*(3), 367–386. DOI: https://doi.org/10.1162/ling.2009.40.3.367

Kush, D., & Dillon, B. (2021). Principle B constrains the processing of cataphora: Evidence for syntactic and discourse predictions. *Journal of Memory and Language*, *120*, 104254. DOI: https://doi.org/10.1016/j.jml.2021.104254

Kush, D., & Eik, R. (2019). Antecedent accessibility and exceptional covariation: Evidence from Norwegian donkey pronouns. *Glossa: a journal of general linguistics*, *4*(1). DOI: https://doi.org/10.5334/gjgl.930

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

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222. DOI: https://doi.org/10.1016/j.jml.2012.02.005

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. DOI: https://doi.org/10.1016/j.jmva.2009.04.008

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

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

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman; Hall/CRC. DOI: https://doi.org/10.1201/9781315372495

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123. DOI: https://doi.org/10.1023/A:1005184709695

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 817–835. DOI: https://doi.org/10.1037/0278-7393.27.3.817

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91. DOI: https://doi.org/10.1016/S0749-596X(02)00515-6

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In D. Luce (Ed.), *Handbook of mathematical psychology*. John Wiley & Sons.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, *2*(1), 28–34. DOI: https://doi.org/10.1007/s42113-018-0019-z

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology*, *7*. DOI: https://doi.org/10.3389/fpsyg.2016.00280

Nicenboim, B., Schad, D., & Vasishth, S. (2021). *An introduction to Bayesian data analysis for cognitive science*. https://vasishth.github.io/Bayes_CogSci/

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34. DOI: https://doi.org/10.1016/j.jml.2017.08.004

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*(S4), 1075–1100. DOI: https://doi.org/10.1111/cogs.12589

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, *6*, 312. DOI: https://doi.org/10.3389/fpsyg.2015.00312

Nouwen, R., Brasoveanu, A., van Eijck, J., & Visser, A. (2016). Dynamic semantics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University.

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

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Romoli, J., Khan, M., Sudo, Y., & Snedeker, J. (2015). Resolving temporary referential ambiguity using presupposed content. In *Experimental perspectives on presuppositions* (pp. 67–87). Springer. DOI: https://doi.org/10.1007/978-3-319-07980-6_3

Romoli, J., & Schwarz, F. (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In *Experimental perspectives on presuppositions* (pp. 215–240). Springer. DOI: https://doi.org/10.1007/978-3-319-07980-6_10

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, *1*(1), 75–116. DOI: https://doi.org/10.1007/BF02342617

Rouder, J. N., Tuerlinckx, F., Speckman, P., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*. DOI: https://doi.org/10.3758/PBR.15.6.1201

Ruys, E. G. (2015). On the anaphoricity of too. *Linguistic Inquiry*, *46*(2), 343–361. DOI: https://doi.org/10.1162/LING_a_00184

Sæbø, K. J. (2004). Conversational contrast and conventional parallel: Topic implicatures and additive presuppositions. *Journal of Semantics*, *21*(2), 199–217. DOI: https://doi.org/10.1093/jos/21.2.199

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). Workflow techniques for the robust use of Bayes factors. DOI: https://doi.org/10.48550/arXiv.2103.08744

Schlenker, P. (2008). Be articulate: A pragmatic theory of presupposition projection. *Theoretical Linguistics*, *34*. DOI: https://doi.org/10.1515/THLI.2008.013

Schmitz, T., Winkowski, J., Hoeks, M., Nouwen, R., & Dotlačil, J. (2024). Semantic accessibility and interference in pronoun resolution. *Glossa Psycholinguistics*, *3*(1). DOI: https://doi.org/10.5070/G60111424

Schwarz, F. (2007). Processing presupposed content. *Journal of Semantics*, *24*(4), 373–416. DOI: https://doi.org/10.1093/jos/ffm011

Schwarz, F. (2015). Presuppositions vs. asserted content in online processing. In *Experimental perspectives on presuppositions* (pp. 89–108). Springer. DOI: https://doi.org/10.1007/978-3-319-07980-6_4

Stan Development Team. (2021). *Stan modeling language users guide and reference manual*. Version 2.27. https://mc-stan.org

Staum Casasanto, L., Hofmeister, P., & Sag, I. A. (2010). Understanding acceptability judgments: Additivity and working memory effects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*. https://escholarship.org/uc/item/08400553

Szabolcsi, A. (2017). Additive presuppositions are derived through activating focus alternatives. In A. Cremers, T. van Gessel, & F. Roelofsen (Eds.), *Proceedings of the 21st Amsterdam colloquium* (pp. 455–464). https://semanticsarchive.net/Archive/jZiM2FhZ/AC2017-Proceedings.pdf

Tesnière, L. (1965). *Éléments de syntaxe structurale* (2nd ed.). Klincksieck.

Tonhauser, J., Beaver, D., Roberts, C., & Simons, M. (2013). Toward a taxonomy of projective content. *Language*, 66–109. DOI: https://doi.org/10.1353/lan.2013.0001

van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, *9*(4), 333–377. DOI: https://doi.org/10.1093/jos/9.4.333

van der Sandt, R., & Geurts, B. (1991). Presupposition, anaphora, and lexical content. In O. Herzog & C.-R. Rollinger (Eds.), *Text understanding in LILOG: Integrating computational linguistics and artificial intelligence. Final report on the IBM Germany LILOG-project* (pp. 259–296). Springer. DOI: https://doi.org/10.1007/3-540-54594-8_65

van der Sandt, R., & Geurts, B. (2001). Too. In R. van Rooy & M. Stokhof (Eds.), *Proceedings of the 13th Amsterdam colloquium* (pp. 180–185). University of Amsterdam.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., et al. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. DOI: https://doi.org/10.3758/s13423-020-01798-5

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. DOI: https://doi.org/10.1037/0278-7393.33.2.407

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

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

Veríssimo, J. (2021). Analysis of rating scales: A pervasive problem in bilingualism research and a solution with Bayesian ordinal models. *Bilingualism: Language and Cognition*, *24*(5), 842–848. DOI: https://doi.org/10.1017/S1366728921000316

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. DOI: https://doi.org/10.1016/j.jml.2009.04.002