Journal of Memory and Language 94 (2017) 316–339

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis

Lena A. Jäger[a,*], Felix Engelmann[b], Shravan Vasishth[a,c]

[a] University of Potsdam, Germany
[b] University of Manchester, UK
[c] Centre de Recherche en Mathématiques de la Décision, CNRS, UMR 7534, Université Paris-Dauphine, PSL Research University, Paris, France

ARTICLE INFO

Article history:
Received 25 August 2015
revision received 17 January 2017

Keywords:
Cue-based retrieval
Syntactic dependency processing
Interference
Bayesian meta-analysis
Agreement
Reflexives

ABSTRACT

We report a comprehensive review of the published reading studies on retrieval interference in reflexive-/reciprocal-antecedent and subject-verb dependencies. We also provide a quantitative random-effects meta-analysis of eyetracking and self-paced reading studies.

We show that the empirical evidence is only partly consistent with cue-based retrieval as implemented in the ACT-R-based model of sentence processing by Lewis and Vasishth (2005) (LV05) and that there are important differences between the reviewed dependency types. In non-agreement subject-verb dependencies, there is evidence for inhibitory interference in configurations where the correct dependent fully matches the retrieval cues. This is consistent with the LV05 cue-based retrieval account. By contrast, in subject-verb agreement as well as in reflexive-/reciprocal-antecedent dependencies, no evidence for inhibitory interference is found in configurations with a fully cue-matching subject/antecedent. In configurations with only a partially cue-matching subject or antecedent, the meta-analysis reveals facilitatory interference in subject-verb agreement and inhibitory interference in reflexives/reciprocals. The former is consistent with the LV05 account, but the latter is not. Moreover, the meta-analysis reveals that (i) interference type (proactive versus retroactive) leads to different effects in the reviewed dependency types and (ii) the prominence of the distractor strongly influences the interference effect.

In sum, the meta-analysis suggests that the LV05 needs important modifications to account for the unexplained interference patterns and the differences between the dependency types. More generally, the meta-analysis provides a quantitative empirical basis for comparing the predictions of competing accounts of retrieval processes in sentence comprehension.

Introduction

Several researchers in sentence comprehension have argued that the formation of dependencies between non-adjacent words relies on a cue-based retrieval mechanism that leads to interference effects (McElree, 2000; Van Dyke & Lewis, 2003; Lewis, Vasishth, & Van Dyke, 2006; Van Dyke & McElree, 2011). For example, in a sentence like *The girl who the man saw laughed*, the dependency between the main clause subject (*girl*) and the main-clause verb (*laughed*) needs to be completed. In order to complete this dependency when reaching the verb, a memory retrieval is initiated for a noun that is the grammatical subject and has an animate referent. The assumption is that so-called retrieval cues, here *subject* and *animate*, allow the cognitive system to seek out the relevant item in memory by direct access. One appeal of this account is that it assumes the same memory access mechanism for language processing that governs recall in general information processing (Watkins & Watkins, 1975; Anderson & Lebiere, 1998; Anderson et al., 2004; McElree, 2006; Ratcliff, 1978; Van Dyke, 2002).

In this paper, we review the empirical evidence presented in the sentence processing literature and synthesize the evidence quantitatively by means of a Bayesian meta-analysis. We then compare the evidence with the predictions of the computationally implemented cue-based retrieval model of Lewis and Vasishth (2005), henceforth LV05.

The LV05 model is based on the general cognitive architecture *Adaptive Control of Thought-Rational* (ACT-R, Anderson & Lebiere, 1998; Anderson et al., 2004).[1] The LV05 model provides quantitative

* Corresponding author.
E-mail address: lena.jaeger@uni-potsdam.de (L.A. Jäger).

[1] The source code of the LV05 model is available from http://www.ling.uni-potsdam.de/~vasishth/code/LewisVasishthModel05.tar.gz.

http://dx.doi.org/10.1016/j.jml.2017.01.004
0749-596X/

predictions of retrieval speed and accuracy by using an incremental parser that relies on associative retrievals which are subject to activation decay and similarity-based interference. The model's quantitative predictions, derived using simulations, have been investigated by carrying out experiments covering a range of syntactic dependency types:

(i) Subject-verb dependencies
   (a) Subject-verb dependencies (other than agreement) in unimpaired populations (Lewis & Vasishth, 2005; Vasishth & Lewis, 2006; Nicenboim, Logačev, Gattei, & Vasishth, 2016) and in aphasic populations (Patil, Hanne, Burchert, Bleser, & Vasishth, 2016).
   (b) Subject-verb agreement dependencies (Dillon, Mishler, Sloggett, & Phillips, 2013; Wagers, Lau, & Phillips, 2009).
(ii) Antecedent-reflexive dependencies (Dillon et al., 2013; Jäger, Engelmann, & Vasishth, 2015; Parker & Phillips, 2014; Patil, Vasishth, & Lewis, 2016) and antecedent-reciprocal dependencies (Kush & Phillips, 2014; Kush, 2013);
(iii) Negative polarity items (Parker & Phillips, 2016; Vasishth, Bruessow, Lewis, & Drenhaus, 2008);
(iv) General dependency resolution difficulty in a large-scale model of parsing (Boston, Hale, Vasishth, & Kliegl, 2011)

In this paper, we will focus on the empirical evidence from the first three types of dependencies (ia, ib and ii) in unimpaired adult native speakers, because evidence from mainly these dependency types has been invoked to argue in favor of or against cue-based memory retrieval subserving sentence processing. The comparison between experiments on interference effects in reflexives and subject-verb agreement has even led researchers to argue that subject-verb number agreement and reflexive-antecedent dependency processing rely on qualitatively different memory access mechanisms (Dillon et al., 2013; Phillips, Wagers, & Lau, 2011). Moreover, the experimental designs used in experiments examining these three types of dependencies are very similar across studies. This makes it possible to quantitatively summarize this literature in a Bayesian meta-analysis.

Target-match and target-mismatch configurations

In this review, we focus on four key syntactic configurations that are often used to investigate effects of retrieval interference in sentence processing. These are shown in Example 1, and are taken from Sturt (2003). We will use this example to introduce key terminology that is used in the present paper; a summary of the terms appears in Table 1. In Example 1, the reflexive *himself* or *herself* must be connected with its antecedent, *surgeon*. Hence, when reading or hearing the reflexive, a retrieval process must be triggered to access the antecedent. We will refer to the noun that is the syntactically correct antecedent (*surgeon*) as the **target** of the retrieval process. The target must be a noun phrase inside the reflexive's binding domain that c-commands the reflexive (Chomsky, 1981). We will say that in this case a **retrieval cue**, *c-command*, is set by the reflexive *himself/herself* to seek out a noun that has the *+c-command* **feature** (here, *surgeon*).[2] In the examples below, the retrieval specification is shown as a set of cues in curly brackets behind the critical word (the reflexive) that triggers retrieval. The feature value associated with a word is represented by the name of the feature prefixed with either a − (absent) or a + (present). Note that only those features that are subject to the experimental manipulation are considered here. For the sake of simplicity, other cues such as *noun phrase* are not considered.

Table 1
Definitions of key terms used in the present paper in connection with cue-based retrieval as implemented in the ACT-R framework and adopted in the Lewis and Vasishth (2005) model.

Term	Definition
Feature	A property of an item in memory Example: The feature *+animate* in the noun *girl*
Retrieval cue	A property used to seek out an item in memory Example: the retrieval cue *animate* is used to seek out the subject of *laughed*
Target	The item that is the syntactically correct target for retrieval
Distractor	An item that is not the syntactically correct target for retrieval
Misretrieval	The retrieval of a distractor rather than the target
Match	A match occurs when a retrieval cue and a feature on an item have the same value
Mismatch	A mismatch occurs when a retrieval cue and a feature on an item do not have the same value
Cue overload	This occurs when a retrieval cue matches the features of two or more items
Fan	The number of items whose features match a retrieval cue
Fan effect	Reduction in activation of items in memory as a result of a fan $\geqslant 2$
Feature overlap	If any two items have an identical feature value, then we have a feature overlap between the two items
Interference	The consequence of a (partial) match of the distractor with the retrieval cues
Inhibitory effect	A slowdown in processing during retrieval
Facilitatory effect	A speedup in processing during retrieval

(1) a. *Target-match; distractor-mismatch*
The surgeon$^{+masc}_{+c-com}$ who treated Jennifer$^{-masc}_{-c-com}$ had pricked himself$\left\{\begin{array}{c} masc \\ c-com \end{array}\right\}$...

b. *Target-match; distractor-match*
The surgeon$^{+masc}_{+c-com}$ who treated Jonathan$^{+masc}_{-c-com}$ had pricked himself$\left\{\begin{array}{c} masc \\ c-com \end{array}\right\}$...

c. *Target-mismatch; distractor-mismatch*
The surgeon$^{-fem}_{+c-com}$ who treated Jonathan$^{-fem}_{-c-com}$ had pricked herself$\left\{\begin{array}{c} fem \\ c-com \end{array}\right\}$...

d. *Target-mismatch; distractor-match*
The surgeon$^{-fem}_{+c-com}$ who treated Jennifer$^{+fem}_{-c-com}$ had pricked herself$\left\{\begin{array}{c} fem \\ c-com \end{array}\right\}$...

When a feature matches a retrieval cue, we will say that there is a **match**; when it doesn't, there is a **mismatch**. Apart from a *c-command* retrieval cue, we will also assume that the reflexive sets a *gender* cue (*masculine* in the case of *himself*, and *feminine* in the case of *herself*) to seek out a noun with the *+masc/+fem* feature. Note that, as in Example 1a and b, when all retrieval cues match the target's features, we have a **full match** between the retrieval cues and the features of the target; we will therefore call Examples 1a and b **target-match** conditions. By contrast, Examples 1c and d will be called **target-mismatch** conditions: the subject of the sentence that c-commands the reflexive is still referred to as target

[2] Note that in contrast to other syntactic (e.g., *case*) or semantic (e.g., *animacy*) features, *c-command* is a relational feature that one item can only have with respect to another item (i.e., no item can be a c-commander per se, but can only be in a c-commanding relation with another syntactic constituent). Thus, keeping track of the *c-command* features of the items in memory is computationally more complex than keeping track of static, i.e., non-relational, features (Kush, 2013). Although in this paper, we will not pursue the distinction between relational and static cues any further, we want to point out that this distinction is an important issue that should be addressed in future research.

because it is the *syntactically* correct antecedent, but the target's features have only a **partial match** with the retrieval cues (the reflexive has feminine marking and the target, *surgeon*, is masculine by default in English).[3]

In these example sentences, there is also a **distractor** noun, *Jonathan* or *Jennifer*. This noun cannot be a legal antecedent to the reflexive because Principle A of the Binding Theory (Chomsky, 1981) requires that an antecedent noun c-command the reflexive—the distractor lies inside a relative clause, preventing it from c-commanding the reflexive. The distractor noun is interesting for investigating interference in sentence comprehension because in Examples 1b and d, the gender cue matches the gender feature on the distractor. In this case, since not all the retrieval cues match the distractor, there is a partial match between the features of the distractor and the retrieval cues. In the LV05 cue-based retrieval framework, the distractor item is a potential retrieval candidate despite its syntactically unlicensed position, because (as discussed later in more detail) partial matches in ACT-R can occasionally lead to the distractor being retrieved.[4]

The predictions of the ACT-R model

We briefly discuss the predictions of the ACT-R model of cue-based retrieval for both target-match and target-mismatch configurations using Example 1; see Fig. 1 for a graphical summary of the predictions. In each case, the retrieval cues lead to a search of items that have two properties, namely being in a c-commanding position and sharing the gender of the reflexive. These properties are the retrieval cues. There are two relevant items in working memory: The c-commanding target and the non-c-commanding distractor. The gender, i.e., the second retrieval-relevant feature, is manipulated on both, the target and the distractor. Thus, the target always matches at least the structural c-command cue, whereas the distractor's features can never fully match the retrieval cues, but may match the gender cue.

In configuration a of Fig. 1, the target's features match both cues with no matching distractor present, i.e., there is a simple, unambiguous match of both cues with the same item's features. Configuration a is usually compared to configuration b, where the distractor item overlaps with the target in the gender feature. The *masculine* cue is now ambiguous or "overloaded" (Watkins & Watkins, 1975), because it matches the features of both items. Because the gender cue does not discriminate the target from the distractor anymore, in the ACT-R architecture the distractor could be erroneously retrieved instead of the target. Thus, the behavior of the model is non-deterministic.

Another consequence of cue overload in ACT-R theory is that, in cases where the target is retrieved rather than the distractor, the retrieval of the target is slower in configuration b than in a. This is explained in terms of the amount of spreading activation that is distributed from the retrieval cues to all matching items. The presence of a partially matching distractor as in b reduces the amount of activation spread from the *masculine* cue to the target in comparison with a because this activation is shared with the distractor. In ACT-R, this is called the *fan effect*, and the number of items associated with the same cue is called the *fan*. Since an item's activation determines its retrieval latency, the target will be retrieved more slowly in configuration b.[5] We will refer to the empirical observation of a processing slowdown due to the presence of a matching distractor as **inhibitory interference**.[6]

We turn next to target-mismatch configurations c and d in Fig. 1 and Example 1c and d. Here, the syntactically licensed antecedent does not match the gender cue (*feminine*). In configuration c, the cues are unambiguous because the distractor does not match any of them. However, there is only a partial match of the target with the retrieval cues. According to ACT-R, due to the partial match, the target has a lower activation in this case than in the single full-match configuration a, leading to a slower retrieval of the target. Configuration d adds a distractor that matches the gender cue, such that, similar to b, there is a partial match; this makes erroneous retrievals possible. However, unlike b, in d the cues are not ambiguous, since each of them matches only a single item. Thus, there is no fan effect. As a consequence, when compared to c, in d no fan-effect driven reduction in activation occurs. Instead, ACT-R predicts a speedup on average in this case, because the availability of two similarly probable retrieval candidates causes shorter mean retrieval latencies, just as in a race process (e.g., Van Gompel, Pickering, & Traxler, 2001; for simulations demonstrating a race process, see Logačev & Vasishth, 2016). We will refer to the empirical observation of a speedup due to a matching distractor as **facilitatory interference**.

Generally speaking, the facilitory processes due to occasional misretrievals of a distractor item and inhibitory processes caused by the fan effect counteract each other in ACT-R. In target-match configurations, the inhibitory interference is predicted to outweigh the facilitatory processes. In target-mismatch configurations, in contrast, the absence of inhibitory processes leads to the domination of the facilitatory processes resulting in a predicted speedup.

To summarize, the LV05 implementation of a cue-based retrieval model of sentence processing predicts *inhibitory interference* in *target-match* configurations and *facilitatory interference* in *target-mismatch* configurations. Although there is considerable evidence for the fan effect affecting dependency processing (Badecker & Straub, 2002; Chen, Jäger, & Vasishth, 2012; Felser, Sato, & Bertenshaw, 2009; Jäger et al., 2015; Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006, 2011) it is still unclear how robust this finding is, as there is also evidence for effects in

[3] In this example, the target-mismatch conditions are not ungrammatical since *surgeon* can also refer to a female doctor. However, it is assumed that at least in the first stage of parsing the reflexive-antecedent dependency, the reader relies heavily on the stereotypical gender, which in the case of *surgeon* is masculine (Sturt, 2003). While many experimental designs in reflexive interference research manipulate the stereotypical gender in order to avoid ungrammatical sentences, there are also several studies in which stimuli with really ungrammatical target-mismatch conditions were tested, for example by using proper names with unambiguous gender as the reflexives' antecedents. In the subject-verb dependencies reviewed in this paper, the target-mismatch conditions are always ungrammatical.

[4] From now on, for convenience, we will say that a retrieval cue matches or (partially) mismatches an item, even though, strictly speaking, a retrieval cue can only match a feature of an item.

[5] Notice that the cue-based retrieval model proposed by McElree, Foraker, and Dyer (2003) explains inhibitory interference in a different way than ACT-R. ACT-R assumes that a competing distractor causes both an increased probability of misretrievals *and* increased retrieval latencies, whereas in the model proposed by McElree et al. (2003), interference is only reflected in a decreased retrieval probability of the target but not in the speed of the retrieval process. McElree's claim that interference affects only the retrieval probability of the target item and not the latency of the retrieval process is based on his observation that in speed-accuracy tradeoff (SAT) experiments, the intercept and the rate parameter (which represent the retrieval speed) are not affected by interference, whereas the asymptote (which represents the retrieval probability of the target) is sensitive to an interference manipulation (McElree, 2000, 2006; McElree et al., 2003). McElree explains the effects observed in reading times with self-paced reading or eyetracking as a by-product of changes in the retrieval probabilities. The idea here is that misretrievals may trigger a repair process that inflates reading times (McElree, 1993). In this paper, we focus on the predictions of cue-based retrieval in the sense of the ACT-R model.

[6] Note that findings of inhibitory interference have also been interpreted as reflecting *encoding interference*. Within a content-addressable framework, not only the retrieval process but also the encoding and maintenance of items in memory can be affected by the items' mutual similarity. E.g., in the content-addressable memory model proposed by Oberauer and Kliegl (2006), the activation level of a memory item decreases as a function of the number of features it shares with other items and as a function of the number of other items it shares features with. Although encoding interference can certainly impact retrieval latency and accuracy, in the present article, our focus is on the ACT-R retrieval account.

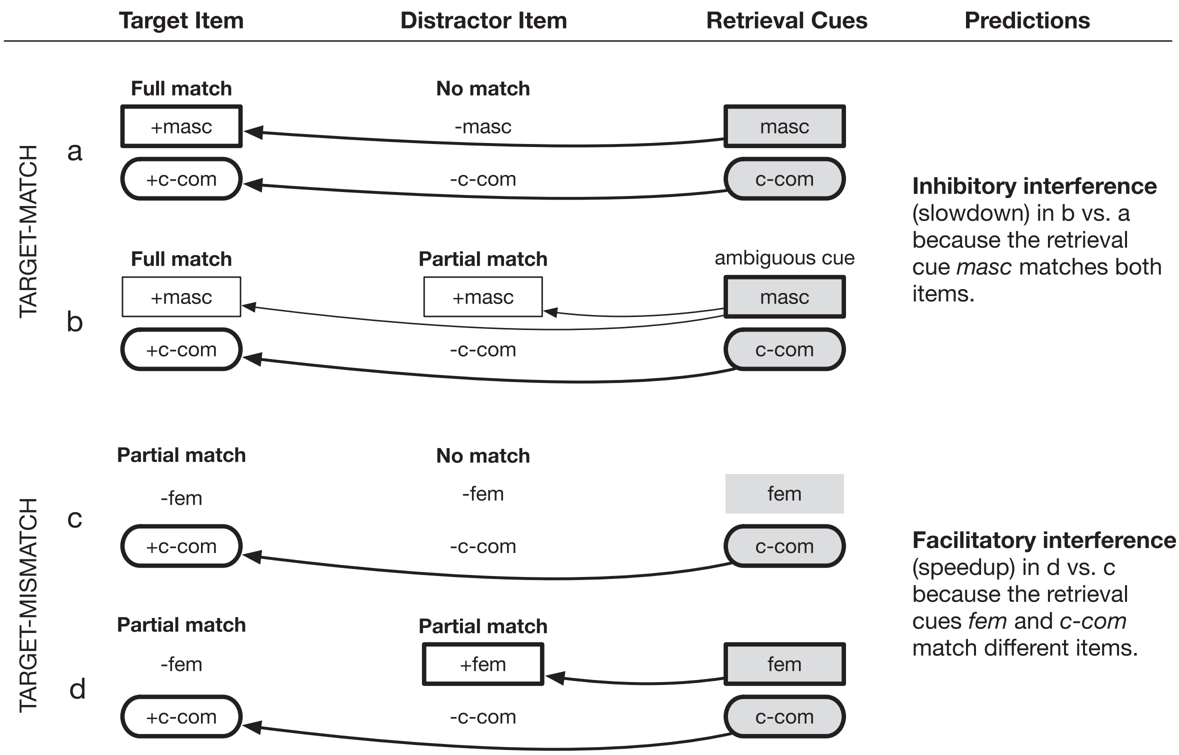


Fig. 1. Predictions of ACT-R for the four conditions shown in Example 1.

the opposite direction (Cunnings & Felser, 2013; Sturt, 2003), as well as a series of statistically non-significant results. Theoretically, it is also under debate whether the parser indeed relies on a cue-based retrieval mechanism and if yes, whether it does so for all kinds of dependencies or only for certain kinds of dependencies (Dillon et al., 2013). As for the predicted facilitation in target-mismatch configurations, the empirical evidence is also mixed. Although some studies show facilitation (Dillon et al., 2013; King, Andrews, & Wagers, 2012; Lago, Shalom, Sigman, Lau, & Phillips, 2015; Parker & Phillips, 2014; Pearlmutter, Garnsey, & Bock, 1999; Tucker, Idrissi, & Almeida, 2015; Wagers et al., 2009), others show inhibition (Cunnings & Felser, 2013; Jäger et al., 2015; Kush & Phillips, 2014).

In sum, the results across the different experiments are mixed. In some experiments, we see facilitatory effects in target-match and inhibitory effects in target-mismatch configurations; these cannot be explained in terms of the ACT-R cue-based retrieval mechanism.[7] The LV05 ACT-R model can only explain inhibitory effects in target-match and facilitation in target-mismatch.

The need to synthesize empirical evidence

It is common in psycholinguistics (e.g., see pp. 155–156 of Phillips et al., 2011) to classify results as statistically significant and non-significant; the implication is that the evidence is in favor of whichever category has more results. This is quite a departure from statistical practice, where conclusions are generally drawn based on estimates of the effect size, and the precision of these estimates. For example, in areas like medicine, meta-analysis—using statistical methods to summarize the results of multiple (independent or dependent) studies (Glass, 1976)—has become an established method for synthesizing evidence as part of a systematic review of the literature (Higgins & Green, 2008). Systematic reviews and meta-analyses, as used in medicine, generally aim to bring together all the available evidence that meets specific criteria in order to make informed decisions about interventions. In the psycholinguistic context, such methods allow us to quantitatively take all the evidence available into account to derive an estimate of the underlying effect of interest. For example, Vasishth, Chen, Li, and Guo (2013) carried out a Bayesian random-effects meta-analysis (Gelman et al., 2014; Sutton, Welton, Cooper, Abrams, & Ades, 2012) on Chinese relative clause data. In the present article, we also carry out a Bayesian meta-analysis to arrive at estimates of the plausible values of the effects of interest. The Bayesian approach is used here for two reasons. First, Bayesian methods provide estimates of plausible values of the parameter of interest; they are thus a powerful tool for directly investigating estimates of the effect given the evidence. Second, with the arrival of probabilistic programming languages like JAGS (Plummer, 2012) and Stan (Stan Development Team, 2013, 2016), it has become very easy to use Markov Chain Monte Carlo methods for sampling from a posterior distribution, which makes it possible to compute estimates for the posterior distribution for essentially any kind of prior and likelihood distributions. For a general review of the application of Bayesian methods in psychology and linguistics, see Nicenboim and Vasishth (2016).

The remainder of this article will present a comprehensive literature review on interference in reflexive-antecedent and subject-verb dependencies, including a Bayesian random effects meta-analysis of the results. We reviewed 110 experimental comparisons on interference in reflexive-antecedent and subject-verb dependencies and performed a meta-analysis on a subset of the reviewed experiments, namely 77 comparisons from reading experiments using self-paced reading or eyetracking published in peer-reviewed journal articles.

In the following, we first present the selection criteria for the studies included in this review and in the meta-analysis, then describe the methodology of the data extraction and of the meta-analysis, and finally present and discuss the results of the

[7] Only under the assumption that the distractor has an extremely high activation as compared to the target does the model predict facilitation in target-match conditions.

qualitative review and the quantitative meta-analysis. The code and data for our meta-analysis are available with the supplementary materials (https://github.com/vasishth/MetaAnalysisJaegerEngelmannVasishth2017).

Inclusion criteria

The experiments included in the Bayesian meta-analysis are summarized in Tables 2 and 3; all the papers that were considered in the qualitative review are listed in the appendix in Tables A1 and A2. We explain the inclusion criteria below. We classify the criteria under the categories: (i) dependency type, (ii) experimental design, (iii) experimental method, and (iv) the sampled population. A detailed documentation as to why certain experiments were not included in the Bayesian meta-analysis is provided in the supplementary materials (see link provided above).

Dependency type

We included experiments investigating interference effects in subject-verb dependencies (agreement and non-agreement dependencies) and in reflexives subject to Binding Principle A (Chomsky, 1981), including reflexives in direct object position, possessive reflexives (e.g., the Chinese *ziji-de* "himself's"), reflexives inside a prepositional phrase (e.g., *of himself*) and reciprocals (e.g., *each other*). Reflexives inside a picture-noun phrase and non-locally bound reflexives were not considered as they presumably differ in their syntactic properties from other types of reflexives (Runner, Sussman, & Tanenhaus, 2003). The materials of the experiments on reflexives and reciprocals were similar to Example 1 discussed above. An exemplary stimulus set of experiments investigating subject-verb number-agreement dependencies taken from Wagers et al. (2009) is provided in Example 2. In this example, *reviewer* is the target and *musician* is the distractor. The feature *local subject* is the structural cue that is always matched by the target and mismatched by the distractor. The interference manipulation is achieved by having the distractor either match or mismatch the *number* cue of the singular verb *praises*. In the target-mismatch conditions, the target mismatches the number of the verb, which results in an ungrammatical sentence.

(2) a. *Target-match; distractor-mismatch*
The musicians$^{-sing}_{-local\ subject}$ who the reviewer$^{+sing}_{+local\ subject}$ praises$\left\{ \begin{array}{l} sing \\ local\ subject \end{array} \right\}$...

b. *Target-match; distractor-match*
The musician$^{+sing}_{-local\ subject}$ who the reviewer$^{+sing}_{+local\ subject}$ praises$\left\{ \begin{array}{l} sing \\ local\ subject \end{array} \right\}$...

c. *Target-mismatch; distractor-mismatch*
The musicians$^{-sing}_{-local\ subject}$ who the reviewers$^{-sing}_{+local\ subject}$ praises$\left\{ \begin{array}{l} sing \\ local\ subject \end{array} \right\}$...

d. *Target-mismatch; distractor-match*
The musician$^{+sing}_{-local\ subject}$ who the reviewers$^{-sing}_{+local\ subject}$ praises$\left\{ \begin{array}{l} sing \\ local\ subject \end{array} \right\}$...

An exemplary item for experiments testing non-agreement subject-verb dependencies is provided in Example 3 taken from Van Dyke (2007, Experiment 1, LoSyn conditions). In this example, the verb *was complaining* triggers the retrieval of its subject. The target of this retrieval process is the subject of the local clause *the resident*. The cue that is manipulated is a semantic one, here the animacy of the target and the distractor (*warehouse* vs. *neighbor*). The underlying assumption here is that the verb *complained* cues for an animate subject.

(3) a. *Target-match; distractor-mismatch*
...the resident$^{+animate}_{+local\ subject}$ who was living near the dangerous warehouse$^{-animate}_{-local\ subject}$ was complaining$\left\{ \begin{array}{l} animate \\ local\ subject \end{array} \right\}$...

b. *Target-match; distractor-match*
...the resident$^{+animate}_{+local\ subject}$ who was living near the dangerous neighbor$^{+animate}_{-local\ subject}$ was complaining$\left\{ \begin{array}{l} animate \\ local\ subject \end{array} \right\}$...

Experimental design

We included experiments that investigated interference effects in either target-match or target-mismatch configurations, where the interference due to the distractor involved any one of several possible retrieval cues (number, gender, etc.). The distractor could be in a proactive or retroactive configuration, and could be within the sentence or presented as memory load before reading the sentence.

Experimental method

We included any comprehension experiment that reports an online processing measure as the dependent variable. In the meta-analysis, we only considered a subset of the reviewed studies (Tables 2 and 3), namely reading experiments using eyetracking or self-paced reading methodology to ensure that the dependent measure was identical (reading time in milliseconds).

Participants

We restricted the review to experiments with linguistically unimpaired, native, adult participants.

Method

Comparisons

We extracted the direction (inhibition or facilitation) and the magnitude of the interference effect within target-match and within target-mismatch conditions. We subtracted the mean of the {*target-match; distractor-mismatch*} condition from the mean of the {*target-match; distractor-match*} condition for the interference effect in target-match configurations. Analogously, we subtracted the mean of the {*target-mismatch; distractor-mismatch*} condition from the mean of the {*target-mismatch; distractor-match*} condition for the interference effect in target-mismatch configurations. A positive effect therefore indicates inhibitory interference, and a negative effect indicates facilitatory interference. An exception is the d' measure of SAT experiments, where a negative effect indicates inhibition.

Note that across publications, the labels of the experimental conditions vary substantially. For example, some authors label the feature overlap between the target and the distractor noun as 'match' rather than the match of the target's/distractor's features

Table 2
Meta-analysis: Mean (first-pass) reading times (ms) of interference effects in experiments on subject-verb dependency comprehension (standard errors in parentheses).

	Publication	Cue	Lang.	Method	Interf. type	Distractor position	Singular verb		Plural verb	
							Target-match	Target-mismatch	Target-match	Target-mismatch
Subject-verb number agreement										
1	Dillon et al. (2013), Exp1 agrmt	num	EN	ET	retro	obj	−14 (16)	–	–	−7 (22)
2	Franck et al. (2015), Exp1 compl	num	FR	SPR	pro	obj	32 (33)	–	–	–
3	Franck et al. (2015), Exp1 RC	num	FR	SPR	pro	obj	110 (48)	–	–	–
4	Lago et al. (2015), Exp1	num	SP	SPR	pro	subj	−4 (14)	–	–	post −40 (14)
5	Lago et al. (2015), Exp2	num	EN	SPR	pro	subj	−7 (8)	–	–	post −36 (18)
6	Lago et al. (2015), Exp3a	num	SP	SPR	pro	subj	post −12 (6)	–	–	post −15 (7)
7	Lago et al. (2015), Exp3b	num	SP	SPR	pro	subj	12 (9)	–	–	post −22 (11)
8	Pearlmutter et al. (1999), Exp1	num	EN	SPR	retro	PP	−35 (10)	–	–	19 (10)
9	Pearlmutter et al. (1999), Exp2	num	EN	ET	retro	PP	post −36 (18)	–	–	post −4 (18)
10	Pearlmutter et al. (1999), Exp3	num	EN	SPR	retro	PP	−36 (10)	–	post 24 (10)	–
11	Tucker et al. (2015)	num	AR	SPR	retro	obj	post −7 (7)	–	–	−29 (14)
12	Wagers et al. (2009), Exp2	num	EN	SPR	pro	subj	−8 (13)	–	–	post −51 (23)
13	Wagers et al. (2009), Exp3	num	EN	SPR	pro	subj	−1 (16)	−33 (23)	13 (17)	post −31 (29)
14	Wagers et al. (2009), Exp4	num	EN	SPR	retro	PP	−27 (13)	–	–	post −42 (17)
15	Wagers et al. (2009), Exp5	num	EN	SPR	retro	PP	post −11 (11)	–	–	post −37 (16)
16	Wagers et al. (2009), Exp6	num	EN	SPR	retro	PP	0 (12)	–	–	–
Non-agreement subject-verb dependencies										
17	Van Dyke (2007), Exp1 LoSyn	sem	EN	SPR	retro	PP	54 (34)	–	–	–
18	Van Dyke (2007), Exp2 LoSyn	sem	EN	ET	retro	PP	post 44 (19)	–	–	–
19	Van Dyke (2007), Exp3 LoSyn	sem	EN	ET	retro	PP	8 (8)	–	–	–
20	Van Dyke and McElree (2006)	sem	EN	SPR	pro	mem	38 (20)	–	–	–
21	Van Dyke and McElree (2011), Exp1b pro	sem	EN	ET	pro	subj	5 (8)	–	–	–
22	Van Dyke and McElree (2011), Exp1b retro	sem	EN	ET	pro	subj	−2 (11)	–	–	–
23	Van Dyke and McElree (2011), Exp2b pro	sem	EN	ET	pro	obj	7 (9)	–	–	–
24	Van Dyke and McElree (2011), Exp2b retro	sem	EN	ET	retro	obj	−7 (9)	–	–	–
25	Van Dyke (2007), Exp1 LoSem	subj	EN	SPR	retro	PP/subj	13 (30)	–	–	–
26	Van Dyke (2007), Exp2 LoSem	subj	EN	ET	retro	PP/subj	37 (21)	–	–	–
27	Van Dyke (2007), Exp3 LoSem	subj	EN	ET	retro	PP/subj	20 (11)	–	–	–
28	Van Dyke and Lewis (2003), Exp4 unambig	subj	EN	SPR	retro	PP/subj	56 (25)	–	–	–

Note. The experiments are ordered by dependency subtype (subject-verb number agreement, non-agreement subject-verb dependencies) and investigated cue (num = number, sem = semantic, subj = subject). The columns Target-Match and Target-Mismatch contain the estimated interference effect within target-match and target-mismatch configurations of singular and plural verbs, respectively, in reading times (self-paced reading experiments) or first-pass reading times (eyetracking experiments) in milliseconds. Positive values indicate inhibition, negative values indicate facilitation. Standard errors are in parentheses. The term 'post' before an effect means that this effect was observed at the post-critical region, otherwise we report the observation at the critical region. The experiments are classified by language (AR = Arabic, EN = English, FR = French, SP = Spanish), experimental method (ET = eyetracking-while-reading, SPR = self-paced reading), interference type (pro = proactive, retro = retroactive), and by syntactic position of the distractor (subj = subject, obj = object, PP = inside a prepositional phrase, mem = three sentence external memory load words). The publication column contains the publication, the experiment number if multiple experiments are reported in one publication, and in case an additional factor was manipulated, the level of this factor.

with the retrieval cues. Moreover, in most of the studies investigating subject-verb agreement dependencies, the authors compare different conditions than we do. This is because the authors were interested in so-called agreement attraction effects, i.e., the attenuation of a grammaticality effect due to the presence of a matching distractor, rather than interference effects. We therefore recoded the comparisons to make them reflect the effects we are interested in. Note that this recoding obviously leads to different estimated effect sizes being presented in our literature review as compared to the effect sizes reported in the respective papers.

Selection of dependent variables, regions of interest, and computation of the estimates

The reviewed studies differ substantially in the coding of the regions of interest. For our review, we were always faithful to the separation of regions as reported in the respective publication. We label the region containing the verb or the reflexive as *critical region*, and as *post-critical* region the region following the critical region, no matter how many words the authors designated as the post-critical region.[8] We did not consider effects in any other region of the sentence. As for the selection of the dependent variables (in eyetracking experiments, usually multiple dependent measures are reported), we first extracted the effects in the measures and regions that were reported as statistically significant by the authors (see Tables A1 and A2 in the appendix). The effects—the sample means of the interference effects in target-match and target-mismatch—reported in these two tables are based on the numbers the authors present in their publications.

Second, we extracted the effects and standard errors that were needed in the meta-analysis (see Tables 2 and 3) in the following way: We extracted the size and the standard error of the interference effect in reaction time (self-paced reading) or first-pass reading time (eyetracking) at the critical region either from the numbers provided in the respective paper or, if available, from the raw data.[9,10] In the latter case, we obtained the estimates of the effects by fitting a linear mixed-effects model with the comparisons described above as fixed effects and a full random effects

[8] The reason we chose to follow the authors' decisions about region of interest was that even when we had the original data, the regions of interest had been fixed by the authors during data preprocessing. We therefore could not align the regions of interest across experiments in a completely consistent way.

[9] Apart from our own data (Chen et al., 2012; Jäger et al., 2015; Patil et al., 2016), we were able to obtain the raw data from the following publications: Cunnings and Felser (2013), Cunnings and Sturt (2014), Dillon et al. (2013), Franck, Colonna, and Rizzi (2015), Kush and Phillips (2014), Lago et al. (2015), Tucker et al. (2015), and Wagers et al. (2009). We are very grateful to these authors for generously releasing their data.

[10] Any discrepancies in the numbers reported in Tables 2 and 3 vs. Tables A1 and A2 are due to the fact that Tables 2 and 3 report numbers extracted from the raw data, and Tables A1 and A2 report the published numbers.

Table 3
Meta-analysis: Mean (first-pass) reading times (ms) of interference effects in experiments on reflexive/reciprocal-antecedent dependency comprehension (standard errors in parentheses).

	Publication	Cue	Lang.	Method	Interf. type	Distractor position	Target-match	Target-mismatch
Reflexives (direct object)								
29	Jäger et al. (2015), Exp1	anim	CN	ET	retro	subj	−3 (5)	22 (7)
30	Jäger et al. (2015), Exp2 local	anim	CN	ET	pro	mem	17 (8)	–
31	Felser et al. (2009), Exp2b natives, inacc-mism	c-com	EN	ET	pro	subj, topic	4 (9)	–
32	Badecker and Straub (2002), Exp3	gend	EN	SPR	pro	subj	post 42 (28)	–
33	Badecker and Straub (2002), Exp5	gend	EN	SPR	pro	gen	2 (13)	–
34	Badecker and Straub (2002), Exp6	gend	EN	SPR	pro	prep. obj	0 (10)	–
35	Cunnings and Felser (2013), Exp1 HI	gend	EN	ET	pro	subj, topic	−2 (14)	−2 (14)
36	Cunnings and Felser (2013), Exp1 LO	gend	EN	ET	pro	subj, topic	−5 (22)	−2 (16)
37	Cunnings and Felser (2013), Exp2 HI	gend	EN	ET	retro	subj, topic	0 (18)	4 (17)
38	Cunnings and Felser (2013), Exp2 LO	gend	EN	ET	retro	subj, topic	−47 (15)	26 (15)
39	Cunnings and Sturt (2014), Exp1	gend	EN	ET	pro	subj, topic	−1 (9)	post 37 (17)
40	Felser et al. (2009), Exp2b natives, no c-com	gend	EN	ET	pro	subj, topic	3 (8)	–
41	Patil et al. (2016)	gend	EN	ET	retro	subj	−13 (18)	10 (12)
42	Sturt (2003), Exp1	gend	EN	ET	pro	subj, topic	−5 (30)	−7 (30)
43	Sturt (2003), Exp2	gend	EN	ET	retro	obj, topic	12 (10)	15 (10)
44	Dillon et al. (2013), Exp1 refl	num	EN	ET	retro	obj	1 (16)	−7 (19)
45	Dillon et al. (2013), Exp2 himself	num	EN	ET	retro	obj	−14 (14)	−10 (14)
46	Dillon et al. (2013), Exp2 themselves	num	EN	ET	retro	obj	−14 (16)	30 (15)
Possessive reflexives								
47	Chen et al. (2012), local	anim	CN	SPR	retro	subj	5 (13)	–
Reciprocals								
48	Badecker and Straub (2002), Exp4	num	EN	SPR	pro	subj	post 48 (37)	–
49	Kush and Phillips (2014)	num	HI	SPR	retro	prep. obj	3 (54)	post 21 (32)

Note. The experiments are ordered by dependency subtype (direct object reflexives, possessive reflexives and reciprocals) and investigated cue (anim = animacy, c-com = c-command, gend = gender, num = number). The columns Target-Match and Target-Mismatch contain the estimated interference effect within target-match and target-mismatch configurations of singular and plural verbs, respectively, in reading times (self-paced reading experiments) or first-pass reading times (eyetracking experiments) in milliseconds. Positive values indicate inhibition, negative values indicate facilitation. Standard errors are in parentheses. The term 'post' before an effect means that this effect was observed at the post-critical region, otherwise we report the observation at the critical region. The experiments are classified by language (CN = Chinese, EN = English, HI = Hindi), experimental method (ET = eyetracking-while-reading, SPR = self-paced reading), interference type (pro = proactive, retro = retroactive), and by syntactic position of the distractor (subj = subject, obj = object, gen = genitive attribute, mem = three sentence external memory load words, topic = discourse topic). The publication column contains the publication, the experiment number if multiple experiments are reported in one publication, working memory capacity (HI = high, LO = low) if applicable, and, in case an additional factor was manipulated, the level of this factor.

structure (without estimating correlations between random intercepts and slopes). Whenever possible, we applied the same data trimming procedure as the authors of the respective publication in order to represent the reported results in the original publication as faithfully as possible.[11] We chose to base the meta-analysis on first-pass reading time as it is the most commonly reported eyetracking measure in the psycholinguistic literature and arguably reflects early cognitive stages of dependency formation (Clifton, Staub, & Rayner, 2007; Vasishth, von der Malsburg, & Engelmann, 2013).[12] In those experiments where the authors observed a significant effect only at the post-critical region, we used the effect at this region for the meta-analysis. A detailed documentation of how we derived the effect sizes and standard errors for each study is provided in the supplementary materials available at https://github.com/vasishth/MetaAnalysisJaegerEngelmannVasishth2017.

Covariates between experiments

The design and methodology of the reviewed experiments varied in several respects that might influence the observed interference effect. When tabulating the data, we therefore included language, experimental method, the retrieval cue under examination, interference type (proactive vs. retroactive) and syntactic position of the distractor as covariates in our review. Those covariates that appeared to affect the direction and magnitude of the interference effect were further investigated in the quantitative meta-analysis (see below). A few studies orthogonally manipulated one of these factors within one experiment in addition to the target-/distractor-match/mismatch manipulations. For example, Van Dyke and McElree (2011) manipulated interference type (pro-/retroactive interference), and within each interference type was a distractor-match/mismatch manipulation, i.e., distractor-match/mismatch was nested within interference type. Because of this nesting structure, we treat the match/mismatch manipulation within interference type as yielding separate data points; as a consequence, when tabulating the data, we include the effects for proactive and retroactive interference as separate lines. Similarly, several studies orthogonally manipulated the match/mismatch of two different retrieval cues. For example, Van Dyke (2007) manipulated the match/mismatch of a semantic cue and a syntactic cue in a fully crossed factorial design. When tabulating the data, we entered the match/mismatch manipulation involving one cue while holding the other cue constant.

Bayesian meta-analysis

We carried out a Bayesian meta-analysis (Sutton et al., 2012; Gelman et al., 2014) of the data considered here. One way to conduct a meta-analysis is to carry out a so-called fixed-effects meta-analysis (Chen & Peace, 2013). This assumes that all the studies have a true effect θ. Thus, if the observed effects from i studies are $\hat{\theta}_i$, then, due to the central limit theorem, for a large enough number of studies, the fixed-effects model is $\theta_i \sim Normal(\theta, \sigma_i^2)$. If, however, it is more reasonable to assume that each study has a different θ, then one can conduct a so-called random-effects meta-analysis. This would assume that each study i has an underlying true mean θ_i that is generated from a normal distribution

[11] Note that in some cases, the data trimming that the authors carried out had a considerable impact on the estimated effect sizes and standard errors.

[12] Early dependent measures are especially relevant for reflexives, which have been argued to show immunity from interference effects at the early stages of processing (Sturt, 2003).

$Normal(\theta, \tau^2)$, and that each observed effect y_i is generated from $Normal(\theta_i, \sigma_i^2)$, where σ_i is the true (unknown) standard error of study i (this is estimated from the standard error s_i in the data). Thus, the random-effects meta-analysis has a new parameter, τ^2, that characterizes between-study variance. The fixed-effects meta-analysis is in fact just a special case of the random-effects model under the assumption that $\tau = 0$.

In this paper, we present a random-effects meta-analysis because it is likely that there is significant heterogeneity in the studies, since they were run under different conditions with different languages, different methods, and in different labs. In this random-effects meta-analysis, we modeled the interference effect within target-match and target-mismatch configurations separately. Moreover, we were interested in quantifying whether the type of interference (proactive as opposed to retroactive interference) and the prominence of the distractor affect the magnitude of the interference effect. In the literature, it has been observed that retroactive interference is stronger as compared to proactive interference (Van Dyke & McElree, 2011). Moreover, in the qualitative literature review, we noticed that interference effects appear to be stronger when the distractor is in a syntactic position that is also prominent (e.g., a subject or a topic). Including interference type and prominence in the Bayesian meta-analysis allows us to evaluate these observations quantitatively. In general, interference type and distractor prominence were not systematically manipulated in the studies but varied across the different experiments. However, their impact on the interference manipulation can be investigated in an exploratory manner within the framework of the Bayesian random-effects meta-analysis by adding them as covariates in the meta-analysis (Berkey, Hoaglin, Mosteller, & Colditz, 1995; Sutton et al., 2012). Each study's estimates for the regression coefficients are treated as being generated from a normal distribution with mean β, where β represents the effect of the factor of interest. We applied sum contrast coding to the interference type factor with +0.5 for proactive interference and -0.5 for retroactive interference, and we coded three levels of distractor prominence based on its position in the sentence: If the distractor is neither a subject nor the topic of the sentence, we coded it as *other*, if the distractor is either a subject or the topic, we coded it as *OR* and if the distractor is both, a subject and the topic, we coded it as *AND*. We applied successive differences coding (Venables & Ripley, 2002), otherwise known as sliding contrasts, to these three levels comparing *other* (coded as -0.5) to *OR* (coded as +0.5) and *OR* (coded as -0.5) to *AND* (coded as +0.5). Interaction terms were not included in the meta-regression because of the sparsity of the data.

As mentioned in the introduction, it has been proposed that the memory access mechanisms may differ by dependency type (Dillon et al., 2013; Phillips et al., 2011). This claim was based on the observation that the sensitivity to interference manipulations varies as a function of dependency type. Therefore, as a second step, we subsetted the data by dependency type (non-agreement subject-verb dependencies; subject-verb number agreement; reflexive-/reciprocal-antecedent dependencies) and repeated the meta-regression for each dependency type separately. Because there was too little data in each subgroup to do a meta-regression with multiple predictors, we only analyzed one predictor in this second analysis: the effect of proactive vs. retroactive interference.

In short, in a first step, we fit one model for target-match and one for target-mismatch configurations for all dependency types together, with the size of the interference effect in each study as dependent variable, and with interference type (pro-/retroactive) and distractor prominence as regression coefficients. Then, in a second step, we modeled the interference effect separately for the three dependency types with interference type as predictor. The model specification was as follows.

Assume that:

- y_i be the observed effect in milliseconds in the i-th study with $i = 1, \ldots, n$.
- θ is the true (unknown) effect, to be estimated by the model.
- σ_i^2 is the true variance of the sampling distribution; each σ_i is estimated from the standard error available from the study i.
- The variance parameter τ^2 represents between-study variance.
- `predictor` is a regression predictor (e.g., proactive vs. retroactive interference sum-contrast coded as +0.5 and -0.5).

We can construct a hierarchical meta-regression model as follows:

$$\begin{aligned}
&y_i|\theta_i, \beta, \sigma_i^2 \sim N(\theta_i + \beta \times \texttt{predictor}_i, \sigma_i^2) \quad i = 1, \ldots, n \\
&\theta_i|\theta, \tau^2 \sim N(\theta, \tau^2), \\
&\theta \sim N(0, 100^2), \\
&\beta \sim N(0, 100^2), \\
&\tau \sim N(0, 100^2), \; \tau > 0 \text{ (truncated normal)}
\end{aligned} \tag{1}$$

θ_i is now the treatment effect in the i-th study adjusted for the predictor effect β. The posterior distribution of θ is also adjusted for β. When more than one predictor is involved (e.g., when examining the effect of prominence as well as pro-/retroactive interference), then the number of β parameters increases.

It is important to remember that causality is difficult to establish here; the analysis is now an observational study and has to be considered exploratory. All the conclusions we present, especially from the meta-regression coefficients, should be seen as tentative and need to be validated in future work using planned experimental designs with high statistical power.

Results of the qualitative analysis

The qualitative analysis complements the quantitative meta-analysis in various respects. First, in the qualitative analysis we can consider a wider range of experiments using different experimental methodologies, including those that did not meet the stricter inclusion criteria of the quantitative meta-analysis. For example, whereas it is not reasonable to compare ERP data and reading times quantitatively, they can be compared qualitatively. Second, the qualitative analysis allows for a discussion of the specific characteristics of each study, whereas the meta-analysis necessarily collapses over these specificities. Although the qualitative analysis should not be used to draw conclusions, it serves to guide the quantitative meta analysis and also future research.

We distinguish between experiments on subject-verb agreement and other subject-verb dependencies because these two groups show different patterns in target-match configurations,[13] which might indicate that the cognitive processes responsible for the observed phenomena are qualitatively different.

Tables A1 and A2 in the appendix provide a comprehensive overview of all the studies considered in this qualitative review (including the studies that did not appear in the Bayesian meta-analysis).

Non-agreement subject-verb dependencies

Experiments investigating non-agreement subject-verb dependencies exclusively focused on the retrieval of its subject by a

[13] In target-mismatch configurations, no non-agreement data are available.

singular verb in target-match configurations. These experiments were only conducted in English and investigated either a purely semantic cue (i.e., a semantic feature that the verb subcategorized its subject for) or the syntactic *subject* feature. These studies consistently report inhibitory interference, which is in line with the predictions of cue-based retrieval (Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006, 2011). Moreover, there is evidence that the inhibitory effect is larger in a retroactive interference configuration than in a proactive one. The strongest evidence for this observation comes from Van Dyke and McElree (2011) who directly compared pro- and retroactive interference within one experiment. The data from another experiment reported by Van Dyke suggest that interference caused by syntactic cues is stronger as compared to interference caused by semantic cues (Van Dyke, 2007). In the meta-analysis, we will quantitatively evaluate the effect of interference type (pro- vs. retroactive) across experiments. The effect of semantic versus syntactic cues, in contrast, cannot be modeled because not enough data are available.

Subject-verb agreement dependencies

For studies on subject-verb number agreement, we distinguish between the processing of singular and plural verbs in this qualitative review.[14] This is because it has been claimed that the agreement process is inherently different when using plural distractors as compared to singular distractors due to the morphological markedness of English plurals compared to the unmarked singular forms (Lau, Rozanova, & Phillips, 2007; Lehtonen, Niska, Wande, Niemi, & Laine, 2006; New, Brysbaert, Segui, Ferrand, & Rastle, 2004).[15]

The experiments on subject-verb agreement exclusively investigated number agreement; we are not aware of any experiment investigating interference effects in, for example, person or gender agreement in adult native speakers. Most of the experiments focused on the processing of singular verbs in target-match (i.e., grammatical) or plural verbs in target-mismatch (i.e., ungrammatical) configurations. Singular verbs in target-mismatch were only tested in one ERP study by Kaan (2002) and in self-paced reading by Wagers et al. (2009, Experiment 3). In fact, these two are the only studies that tested all four configurations (target-match/mismatch in singular/plural verbs) within the same experiment.

Target-match configurations

In target-match configurations with singular verbs, almost exclusively facilitatory interference has been found across languages and experimental methods. As the only exception to the pattern of facilitatory effects, Franck et al. (2015, Experiment 1; relative clause constructions) report inhibitory interference in target-match conditions of singular verbs in a self-paced reading experiment in French with proactive interference from a distractor that is the head of a relative clause. Moreover, comparing the mere number of studies reporting non-significant results in singular target-match configurations with the number of studies observing significant effects suggests that facilitatory interference is stronger in the retroactive interference configuration.

It is unclear whether the strength of the interference effect in sentence comprehension is affected by the syntactic position of the distractor, as it has been shown to be the case in production (Hartsuiker, Antón-Méndez, & van Zee, 2001). Both questions will be evaluated quantitatively by including interference type and distractor prominence as covariates in the Bayesian meta-analysis.

The available data for plural verbs in target-match configurations are very sparse and inconclusive. Acuña-Fariña, Meseguer, and Carreiras (2014) as well as Pearlmutter (2000, Experiment 2) report facilitation at the verb, whereas Pearlmutter et al. (1999, Experiment 3) observed inhibitory interference at the post-verbal region. All three experiments tested a retroactive interference configuration with the distractor being contained in a prepositional phrase (PP). In Pearlmutter (2000), this PP contained two distractors, but the facilitatory effect was solely caused by the first one.

Target-mismatch configurations

In target-mismatch conditions of singular verbs, the data are extremely sparse: Kaan (2002) reported an increased positivity in event-related potentials in Dutch materials. In the only other study testing this configuration (Wagers et al., 2009, Experiment 3), no significant effects were observed in eyetracking.

In target-mismatch conditions of plural verbs, facilitatory interference was observed in English (Dillon et al., 2013; Pearlmutter et al., 1999; Wagers et al., 2009), Spanish (Lago et al., 2015), and Arabic (Tucker et al., 2015) in both pro- and retroactive interference configurations. In contrast to the facilitation observed in target-match conditions of singular verbs, the locus of the effect in target-mismatch plural conditions appeared to be somewhat delayed in general: Facilitation was observed at the critical region containing the verb in only three experiments, whereas in eight experiments, the effect reached significance only at the post-critical region. Inhibitory interference was only observed by Pearlmutter et al. (1999, Experiment 1). However, this inhibitory effect turned into a statistically significant facilitation at the post-verbal region. Severens, Jansma, and Hartsuiker (2008) tested Dutch materials in singular-verb target-match and plural-verb target-mismatch conditions using event-related potentials. However, the interference effect with respect to our comparison coding cannot be derived from the data provided in their paper, since distractor-match and distractor-mismatch conditions are analyzed at different time windows and electrodes.

Potential confounds

Wagers et al. (2009) have claimed that in terms of number attraction, an effect is rather expected when using plural distractors compared to singular distractors. This difference, the so-called *number asymmetry*, has been attributed to the morphological markedness of English plurals compared to the unmarked singular forms.

Moreover, in most of the studies reporting interference in singular target-match conditions of subject-verb agreement dependencies, the distractor noun was dominated by a prepositional phrase retroactively interfering with the subject. Wagers et al. (2009) have noted that any facilitatory effect observed in these cases, especially where distractor and verb are adjacent, potentially represent spillover effects from the number manipulation of the distractor, that is, reduced processing difficulty for the shorter and morphologically less complex singular distractor may explain a speed-up in reading times also at the subsequent regions. However, three studies show that number-matching distractors in other syntactic positions also induce interference effects: Nicol, Forster, and Veres (1997) observed retroactive facilitatory interference from a relative clause object in their Experiment 5, Lago et al. (2015) report in Experiment 3a *proactive* facilitatory interference from a distractor which is in the subject position of the matrix clause, and Franck et al. (2015) found proactive *inhibitory* interference from a distractor that is the head of a relative clause. In target-match conditions with a plural verb, the effect of plural complexity would be different: In contrast to singular verbs, the predicted number effect for plural verbs is inhibitory, because here the distractor is plural in the distractor-match condition. However,

[14] In the quantitative meta-analysis, this distinction is not possible because not enough data are available.

[15] This distinction is not made for reflexives, as in (English) reflexives the singular/plural distinction is a lexical rather than a purely morphological one.

as mentioned above, the available data for plural target-match conditions are sparse and inconclusive.

In target-mismatch configurations, in contrast, the structural position of the distractor does not seem to have as much of an influence as in the target-match conditions.

In sum, distractor position and plural marking might have confounded several of the experiments on subject-verb agreement. However, this issue is currently impossible to quantitatively address in the meta-analysis as not enough data are available.

Reflexive-/reciprocal-antecedent dependencies

Target-match configurations

Among the studies that did observe significant effects in target-match configurations, most found inhibitory interference (Jäger et al., 2015, Experiment 2; Felser et al., 2009, c-command cue; Badecker & Straub, 2002, Experiments 3 and 4; Patil et al., 2016; Clackson & Heyer, 2014). By contrast, facilitatory interference was found in Sturt (2003, Experiment 1), and in Cunnings and Felser (2013, Experiment 2, participants with low working memory capacity). The majority of the studies, however, report statistically inconclusive results (Jäger et al., 2015, Experiment 1; Badecker & Straub, 2002, Experiments 5 and 6; Cunnings & Felser, 2013, Experiment 1, Experiment 2, high capacity readers; Cunnings & Sturt, 2014; Felser et al., 2009, gender cue; King et al., 2012; Nicol, 1988; Sturt, 2003, Experiment 2; Dillon et al., 2013, Experiment 1; Parker & Phillips, 2014, Experiments 1, 2, 3; Chen et al., 2012; Clackson, Felser, & Clahsen, 2011; Kush & Phillips, 2014).

Target-mismatch configurations

In target-mismatch configurations, statistically significant facilitatory interference was observed in a series of experiments by Parker and Phillips (2014) and also by King et al. (2012). A significant inhibitory interference effect was only observed by Jäger et al. (2015, Experiment 1) in Chinese reflexives. Marginal inhibitory effects were found in Hindi reciprocals by Kush and Phillips (2014), and in English reflexives by Cunnings and Felser (2013, Experiment 2, participants with low working memory capacity) and Cunnings and Sturt (2014, Experiment 1).

In sum, the literature on interference in reflexive/reciprocal processing shows a lot of variability and a considerable number of null results. Moreover, although several studies tested both target-match and target-mismatch configurations, not a single study reports significant results in both configurations. From the published results, it seems that an effect is more likely to be present in target-mismatch than in target-match configurations. This observation has been referred to as *grammatical asymmetry* (Wagers et al., 2009). Although this might just be noise resulting from the generally low statistical power of the studies, it should be noted that even in Jäger et al. (2015, Experiment 1), who tested an unusually large sample size (150 participants), no effect was observed in target-match configurations. Moreover, the pattern of the published results indicates that it is more likely to observe an interference effect in materials where the distractor is in a prominent position, i.e., when it is a subject and/or the discourse topic. Direct evidence for the impact of distractor prominence on the interference effect comes from Parker and Phillips (2014), who manipulated the distractor's prominence relative to the target by varying the target's match with either one or two retrieval cues. In several eyetracking experiments, they compared interference in reflexive-antecedent dependencies in the usual target-mismatch conditions with conditions where the target had *two* mismatching features. In the two-feature mismatch conditions they did observe interference, whereas no effect was detected in the one-feature mismatch conditions. As for interference type, there are no indications in the published experiments that pro- or retroactive configurations would affect the interference effect.

Summary

The qualitative analysis suggests the following patterns of effects.

(1) Non-agreement subject-verb dependencies
- (i) *Target-match.* Inhibitory interference.
- (ii) *Target-mismatch.* No data.

(2) Subject-verb agreement dependencies
- (i) *Target-match*
  - (a) *Singular verb.* Facilitatory interference with the exception of Franck et al. (2015, Experiment 1, relative clause constructions).
  - (b) *Plural verb.* Very sparse data. Facilitatory interference was observed in Acuña-Fariña et al. (2014) and Pearlmutter (2000, Experiment 2), whereas Pearlmutter et al. (1999, Experiment 3) report inhibition.
- (ii) *Target-mismatch*
  - (a) *Singular verb.* Extremely sparse data. Kaan (2002) reports an increased positivity in ERPs.
  - (b) *Plural verb.* Facilitatory interference across languages (English, Spanish, Arabic).

(3) Reflexive-/reciprocal-antecedent dependencies
- (i) *Target-match.* Inhibitory interference with the exception of Sturt (2003, Experiment 1) and Cunnings and Felser (2013, Experiment 2, participants with low working memory capacity).
- (ii) *Target-mismatch.* Several experiments show facilitatory interference; one experiment shows significant inhibitory interference in Chinese (Jäger et al., 2015), and three experiments show marginally significant inhibitory effects in Hindi (Kush & Phillips, 2014) and English (Cunnings & Felser, 2013; Cunnings & Sturt, 2014).

Results of the Bayesian meta-analysis

It may be worth explaining briefly how one can interpret the results of a Bayesian data analysis (for an extended tutorial discussion, see Nicenboim & Vasishth, 2016). Bayesian data analysis uses the data and prior distributions defined on the model parameters to compute the posterior distribution of each parameter. This posterior distribution gives us direct information about plausible values of the model parameters given the data; as a consequence, it becomes possible to make statements about the probability that a parameter is positive (or negative, or in any specific interval), and to calculate a credible interval, i.e., an interval over which we can be 95% certain that the true value of the parameter lies given the data.[16] Hypothesis testing only allows us to decide whether the null hypothesis (which usually says "there is no effect of the experimental manipulation") can be rejected or not given the data. It tells us nothing about the probability of our specific hypothesis of interest being true. By contrast, Bayesian data analysis allows us to quantify our uncertainty about the estimate of the effect of interest. Binary decisions about statistical (non-)significance are no longer at issue. For example, the posterior distribution of the

[16] Cf. the 95% confidence interval, which has a very convoluted meaning: if we were to take many random samples and form a confidence interval from each one, then 95% of these intervals would contain the true population parameter; see Morey, Hoekstra, Rouder, Lee, and Wagenmakers, 2015.

interference effect is a probability distribution that represents the range of plausible values of the effect given the data. In our meta-analysis, we are interested in the probability of there being an inhibitory or a facilitatory effect given the data from the reviewed experiments. That is, we want to compute the probability of the effect being greater than zero (inhibition).[17] This can be done by computing the area under the curve of the posterior distribution associated with a positive effect. The posterior distributions of the regression covariates (interference type and distractor prominence) are interpreted in an analogous fashion. The posterior distribution of interference type quantifies our belief about the effect that a proactive interferer has compared to a retroactive one. For example, assuming that we code proactive interference as +0.5 and retroactive as −0.5, when most of the probability mass of the posterior distribution of interference type lies to the right of zero, this means that a proactive distractor shifts the interference effect to the right, i.e., making it 'more inhibitory' or 'less facilitatory'.

A summary of the data that was used in the quantitative meta-analysis is presented in Tables 2 and 3.

An overview of the results of all random-effects meta regression models is provided in Table 4. For each model, the mean of the posterior estimate of the interference effect and of the regression coefficients (interference type and distractor prominence, if applicable) are presented together with a 95% credible interval, and the posterior probability of the effect being greater than 0. Note that this posterior probability is not a frequentist *p*-value; there is no notion of a "significant" or "non-significant" effect. The probability only serves to quantify our uncertainty about the direction of the effect.

The results of the random-effects meta-regression modeling the interference effect across all dependency types are visualized in Fig. 2 for target-match configurations and in Fig. 3 for target-mismatch configurations. The two figures show the data (mean and 95% confidence interval) together with the posterior estimates (mean and 95% credible interval) of the interference effect for each experiment. Fig. 4 shows the posterior distributions of the interference effect together with the posterior distributions of the effects of interference type (proactive vs. retroactive) and distractor prominence (baseline vs. subject or topic; subject or topic vs. subject and topic) that were obtained in the random-effects meta-regressions that modeled all dependency types together for target-match and target-mismatch configurations, respectively.

The results of the meta-regressions by dependency type are presented in Fig. 5. For each dependency type and target type (match and mismatch), the figure shows the posterior distribution of the interference effect and the posterior distribution of interference type. Target-mismatch configurations are only presented for subject-verb agreement and reflexive-/reciprocal-antecedent dependencies since no data are available for target-mismatch configurations in non-agreement subject-verb dependencies.

Analysis collapsing over dependency types

The analysis of all dependency types together reveals that, when taking interference type and distractor position into account, overall there is no evidence for an interference effect in target-match configurations as the posterior distribution is approximately centered around zero. The posterior estimate for the interference effect is −1.2 ms and the probability of the interference effect being greater than zero is 0.34.[18] In target-mismatch configurations, there is weak evidence for facilitatory interference, the posterior estimate for the interference effect is −5.2 ms and the probability of the effect being facilitatory is 0.87.[19]

Interference type (proactive vs. retroactive) appears to have a strong impact on the interference effect. In target-match configurations, proactive interference leads to a shift of the interference effect in the positive direction (right-shift), i.e., in case of inhibitory interference, it leads to a larger effect size and in case of facilitatory interference, proactive interference leads to a smaller effect size as compared to retroactive interference. In target-mismatch configurations, the pattern is reversed: proactive interference leads to a left-shift of the interference effect, i.e., inhibitory interference effects are smaller and facilitatory effects have a larger effect size in proactive as compared to retroactive configurations.

The prominence of the distractor also appears to have an impact on the interference effect. In target-match configurations, although there is no evidence that being a subject *or* topic (as compared to not being either) affects the strength of the interference, there is some evidence that a distractor that is *both* a subject *and* the topic of the discourse leads to a left-shift of the interference effect as compared to a distractor that is either a subject or a topic. In other words, a very prominent distractor leads to a larger interference effect size in case of facilitation and to a smaller effect size in case of inhibitory interference, or even turns an inhibitory effect into a facilitatory one. In target-mismatch configurations, the effect of distractor prominence is not only more robust, but, critically, goes in the opposite direction as compared to target-match configurations. Both the *other* versus OR and the OR versus AND comparison reveal that a more prominent distractor leads to a more positive interference effect, i.e., larger inhibitory effects or smaller facilitatory effects that possibly turn into inhibitory ones.

Analyses by dependency type

The by-dependency-type analysis shows that the patterns observed in the meta-regression collapsing over dependencies are driven by differences between the dependency types. It reveals that in target-match configurations, there is strong evidence for inhibitory interference in non-agreement subject-verb dependencies: The posterior estimate for the interference effect size is 13.1 ms and the posterior probability of seeing a positive interference effect is 0.99. In experiments on subject-verb agreement, in contrast, there is some evidence for a facilitatory effect with the posterior estimate for the interference effect size being −6.6 ms and the posterior probability of observing a facilitatory effect being 0.91.[20] In reflexive/reciprocal-antecedent dependencies, there is no evidence for interference in target-match configurations as the respective posterior distribution is centered around zero. Hence, the absence of evidence for interference in target-match configurations in the analysis on all data is driven by the opposing patterns in subject-verb agreement and non-agreement dependencies and the absence of an effect in reflexives/reciprocals.

[17] The probability of the effect being smaller than zero (facilitation) is simply calculated by computing 1 minus the probability of the effect being larger than zero.

[18] Modeling the interference effect without interference type and distractor prominence as covariates does not yield any evidence for interference in target-match configurations either; the mean of the interference effect's posterior distribution is 0.1 ms and the probability of a positive interference effect is 0.51.

[19] For target-mismatch configurations, modeling the interference effect without interference type and distractor prominence as covariates yields similar results as the analysis that includes the covariates. The mean of the interference effect's posterior distribution is −5.8 ms and the probability of a negative (facilitatory) interference effect is 0.87.

[20] Concerning the experiments on subject-verb agreement, the qualitative literature review had revealed that Franck et al. (2015, Experiment 1, relative clause constructions) and Pearlmutter et al. (1999, Experiment 1) are the only experiments in which statistically significant *inhibitory* interference was observed. Facilitatory effects are reported more often. The effect reported by Franck et al. (2015) is much larger than any other effect reported in the literature on interference. We have therefore repeated the analysis without the Franck et al. (2015) data. The posterior estimate of the interference effect size changes from −6.6 ms to −7.7 ms, and the posterior probability of observing a facilitatory effect changes from 0.91 to 0.95.

Table 4
Summary of results of the meta-analysis. The symbol b refers to the estimate of the effect of interest. Shown are the mean of the estimated effect $\bar{b}$ (i.e., the mean of the posterior distribution), a 95% credible interval, and the posterior probability of the effect being inhibitory (i.e., greater than 0). A positive interference effect means inhibition, a negative one facilitation. A positive effect of interference type means that proactive interference leads to a right-shift of the interference effect (increased inhibition or reduced facilitation) and retroactive interference to a left-shift. A positive effect of the distractor prominence comparisons means that a more prominent distractor leads to a right-shift of the interference effect (i.e., increased inhibition or reduced facilitation).

Dependency	Effect	Target	$\bar{b}$	95% CrI	$P(b > 0)$
All	Interference	Match	−1.2	[−7, 4.9]	0.34
		Mismatch	−5.2	[−14.5, 3.8]	0.13
	Int. type (pro/retro)	Match	8.2	[−3, 19.6]	0.93
		Mismatch	−32.8	[−54.1, −11.1]	0
	Prominence (OR/other)	Match	−2.2	[−17.9, 14]	0.39
		Mismatch	34.9	[5, 63.6]	0.99
	Prominence (AND/OR)	Match	−12.4	[−32.6, 7.4]	0.1
		Mismatch	39.3	[10.8, 67.9]	1
Subject-verb (non-agreement)	Interference	Match	13.1	[1.7, 28.1]	0.99
	Int. type (pro/retro)	Match	−10.2	[−37.3, 12.9]	0.19
Subject-verb agreement	Interference	Match	−6.6	[−16.2, 3.7]	0.09
		Mismatch	−21.9	[−36.4, −9]	0
	Int. type (pro/retro)	Match	16	[−2.9, 36.5]	0.95
		Mismatch	−14.4	[−41.9, 12]	0.13
Reflexive-/reciprocal-antecedent	Interference	Match	0.1	[−6.3, 6]	0.53
		Mismatch	10.9	[−0.7, 22.2]	0.97
	Int. type (pro/retro)	Match	10.3	[−1.7, 23]	0.95
		Mismatch	−6	[−28.6, 17.5]	0.29

In target-mismatch configurations, the by-dependency-type analysis shows that the weak evidence for facilitation is a result of relatively robust opposing patterns in subject-verb agreement and reflexive-/reciprocal-antecedent dependencies. These two opposing effects apparently cancel each other out: (i) facilitatory interference in subject-verb agreement (−21.9 ms, with posterior probability of a facilitatory effect of 1), and (ii) inhibitory interference in reflexive-/reciprocal-antecedent dependencies (10.9 ms, with posterior probability of observing an inhibitory effect of 0.97). Importantly, the posterior estimate indicating inhibitory interference in reflexives and reciprocals is not simply due to the relatively high power experiment on Chinese reflexives presented by Jäger et al. (2015), which is the only study that reports significant inhibitory interference in target-mismatch configurations in reflexives. Removing Jäger et al.'s Experiment 1 from the meta-regression yields a posterior estimate for the interference effect size of 9.5 ms and a posterior probability of the effect being inhibitory of 0.94.

Moreover, the by-dependency-type analyses reveals that interference type (pro- vs. retroactive interference) affects the interference effect in different ways in non-agreement subject-verb dependencies as opposed to the other two dependency types. The right-shift of the interference effect due to proactive interference in target-match configurations in the overall analysis is entirely driven by subject-verb agreement and reflexive-/reciprocal-antecedent dependencies. In non-agreement subject-verb dependencies, in contrast, the interference effect is somewhat shifted to the left in proactive interference configurations. In other words, there is (at least weak) evidence that the inhibitory interference in non-agreement subject-verb dependencies is larger in retroactive interference configurations, whereas the opposite is the case for subject-verb agreement and reflexive-/reciprocal-antecedent dependencies.

In target-mismatch conditions, the effect of interference type in reflexives/reciprocals and in subject-verb agreement is consistent with the results of the analysis that is collapsed over dependency types: proactive interference negatively affects the size of the interference effect (i.e., leads to a left-shift of the interference effect). However, the effects are very weak in the by-dependency-type analysis, whereas the effect is strong in the overall analysis. This discrepancy is probably due to the lower statistical power of the by-dependency-type analysis and, as a more detailed inspection of the models reveals, due to the distractor prominence factor that is taken into account in the model on all data but not in the by-dependency-type analysis (due to sparsity of data).

Discussion

We first discuss the results of our analyses against the background of the LV05 ACT-R model of cue-based retrieval. Then, we discuss methodological issues related to our analyses. Regarding the methodological issues, we first discuss possible reasons for the discrepancies between the results of the qualitative literature review and the Bayesian meta-analysis, and then discuss some of the limitations of our quantitative meta-analysis.

Discussion of the results

Table 5 summarizes the findings of the meta-analysis and compares them to the predictions of the LV05 model.

Non-agreement subject-verb dependencies

The inhibitory interference in non-agreement subject-verb dependencies found in the meta-analysis confirms the conclusions drawn from the qualitative literature review and is in line with the predictions of the LV05 model. As no data are currently available for target-mismatch configurations, an open question for future research is whether the interference pattern in this configuration is also consistent with the LV05 model's predictions.

Subject-verb agreement dependencies

In subject-verb number agreement, the evidence is only partially consistent with cue-based retrieval. In target-match configurations, the meta-analysis shows some evidence for facilitatory interference, which is in line with the conclusions from the qualitative review but stands in contrast to the predicted inhibition of the LV05 model.

In target-mismatch configurations of subject-verb number agreement, the quantitative meta-analysis shows that there is

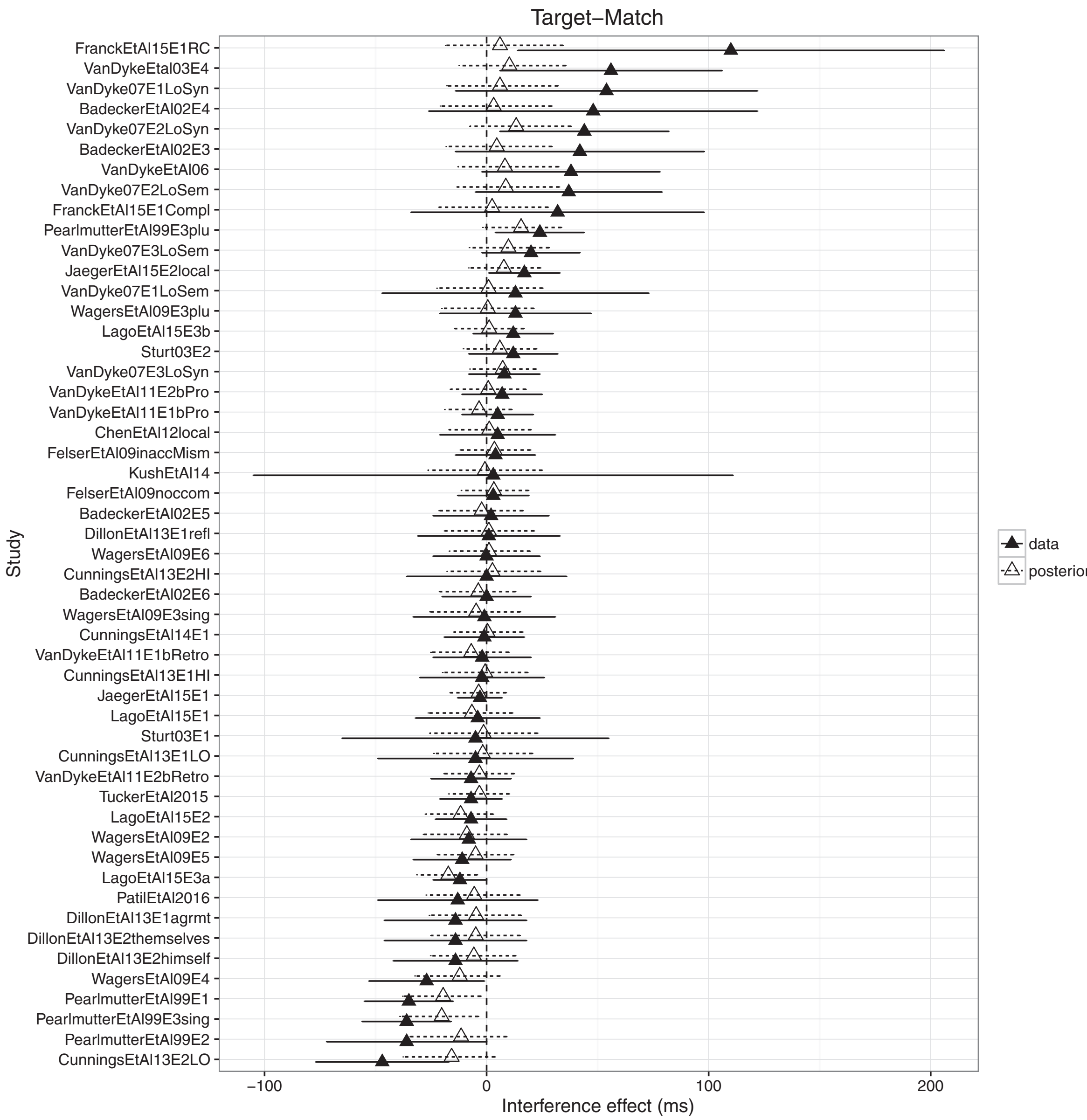


Fig. 2. Summary of the random-effects meta-regression modeling target-match configurations across all dependency types. For each experiment, the figure shows the interference effect in the data (mean and 95% confidence interval) together with the posterior estimate (mean and 95% credible interval) for the interference effect in the respective experiment. A positive effect means inhibition, a negative one means facilitation.

strong evidence for facilitatory interference, confirming the conclusions from the qualitative literature review. This facilitation is predicted by the LV05 model.

Reflexive-/reciprocal-antecedent dependencies

In reflexive- and reciprocal-antecedent dependencies, the results are not compatible with the LV05 model. In target-match configurations, the meta-analysis does not confirm the pattern observed in the qualitative literature review: although several studies report significant inhibitory interference, the posterior distribution of the interference effect obtained in the meta-analysis is centered around zero, meaning that there is no evidence for interference effects affecting the processing of reflexives in target-match configurations. Wagers et al. (2009) noticed that in reflexives, interference manipulations turn out statistically significant more often in target-mismatch as compared to target-match configurations. In order to explain this observation, they proposed that a fully cue-matching target is less prone to interference from a distractor than an only partially cue-matching target. This proposal might explain the absence of an interference effect in target-match configurations within the context of cue-based retrieval. But note that the LV05 model does not assume that the activation of the target affects its sensitivity to inhibitory interference caused by cue-overload (see section '*Distractor prominence*' below).

In target-mismatch configurations of reflexive-/reciprocal-antecedent dependencies, the quantitative modeling shows evidence for inhibitory interference. This is inconsistent with the qualitative literature review: among all studies considered, there is only a single experiment which shows a statistically significant inhibitory effect whereas facilitation is observed in several studies and the vast majority of the experiments show null results. We will discuss reasons for the discrepancy between the quantitative results of the meta-analysis and the qualitative overview below in the section '*Discrepancies between the qualitative review and the quantitative analysis*'. The inhibitory effect revealed by the

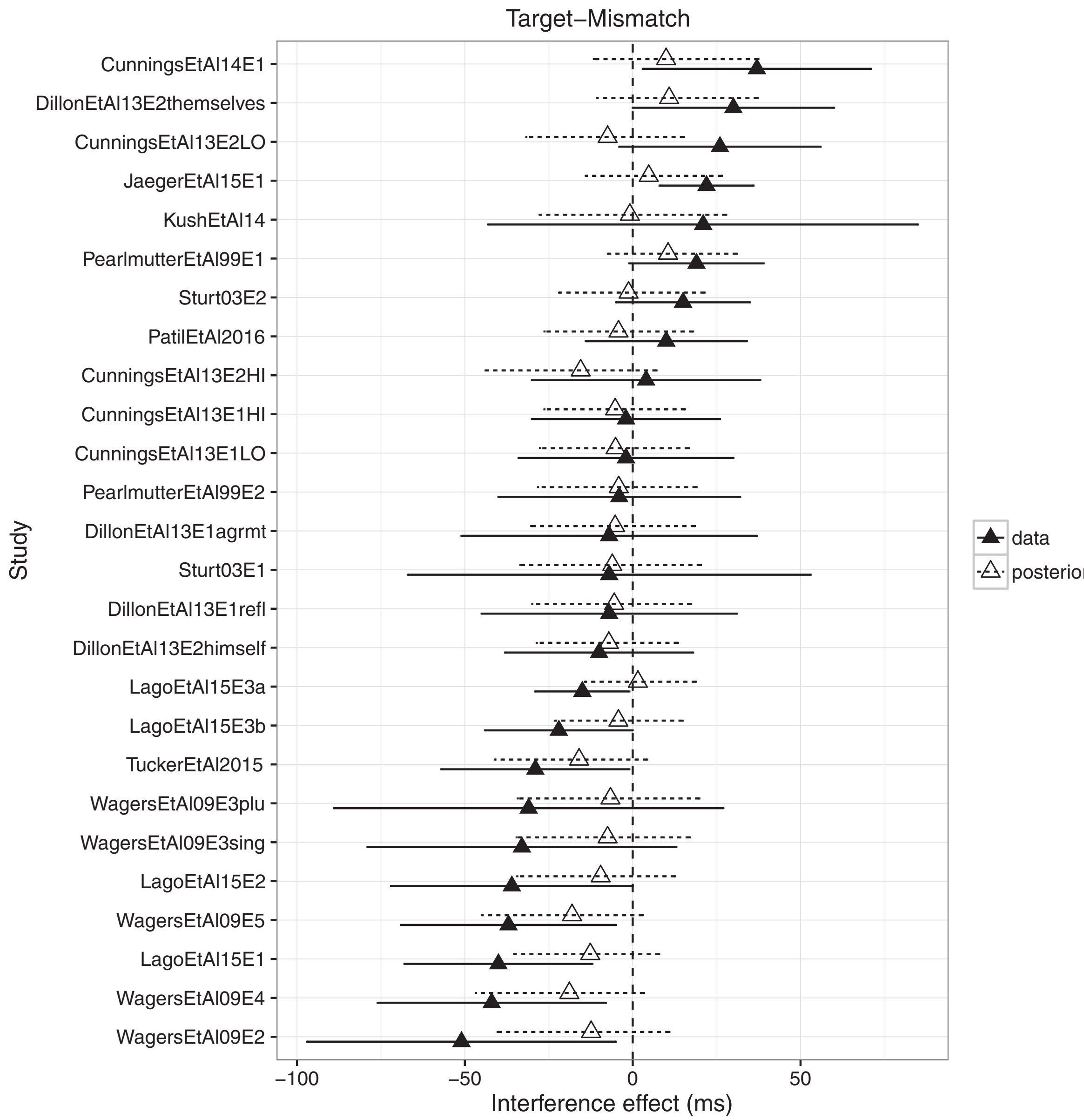


Fig. 3. Summary of the random-effects meta-regression modeling target-mismatch configurations across all dependency types. For each experiment, the figure shows the interference effect in the data (mean and 95% confidence interval) together with the posterior estimate (mean and 95% credible interval) for the interference effect in the respective experiment. A positive effect means inhibition, a negative one means facilitation.

meta-analysis is not compatible with the LV05 model of cue-based retrieval, as the LV05 model predicts facilitation in target-mismatch configurations.

Distractor prominence

The meta-analysis suggests that the distractor's prominence affects the interference effect. The model fitted on all dependency types together reveals that in target-match configurations, a very prominent distractor leads to a left-shift of the posterior distribution of the interference effect; that is, it increases a facilitatory effect or decreases an inhibitory one, or makes an inhibitory effect turn into a facilitatory one. This pattern is consistent with the LV05 model of cue-based retrieval. This is because, generally speaking, in ACT-R, facilitatory processes caused by occasional misretrievals of a distractor and inhibitory similarity-based interference counteract each other. Importantly, a distractor with a higher base level activation (i.e., a more prominent distractor) is misretrieved more often than a less prominent one leading to an increase in facilitation on average across the trials of an experiment. The amount of inhibitory interference, in contrast, is unaffected by the prominence of the distractor in the LV05 ACT-R model.[21] Thus, a more prominent distractor increases the amount of facilitation while leaving the amount of inhibition unaffected. Hence, the empirically observed left-shift of the interference effect (i.e., less inhibition or more facilitation) induced by a more prominent distractor is consistent with the LV05 model.

In target-mismatch configurations, the meta-analysis shows that a more prominent distractor leads to a right-shift of the interference effect, that is, it increases the amount of *inhibitory interference*. However, in ACT-R a more prominent distractor should lead to more misretrievals and therefore to an increase in *facilitatory interference* as in the target-match configurations. Only in the case of a distractor with a much higher activation than the target is the

[21] In ACT-R, the amount of inhibitory interference associated with a retrieval cue is solely determined by the *number* of distractors that match this cue.

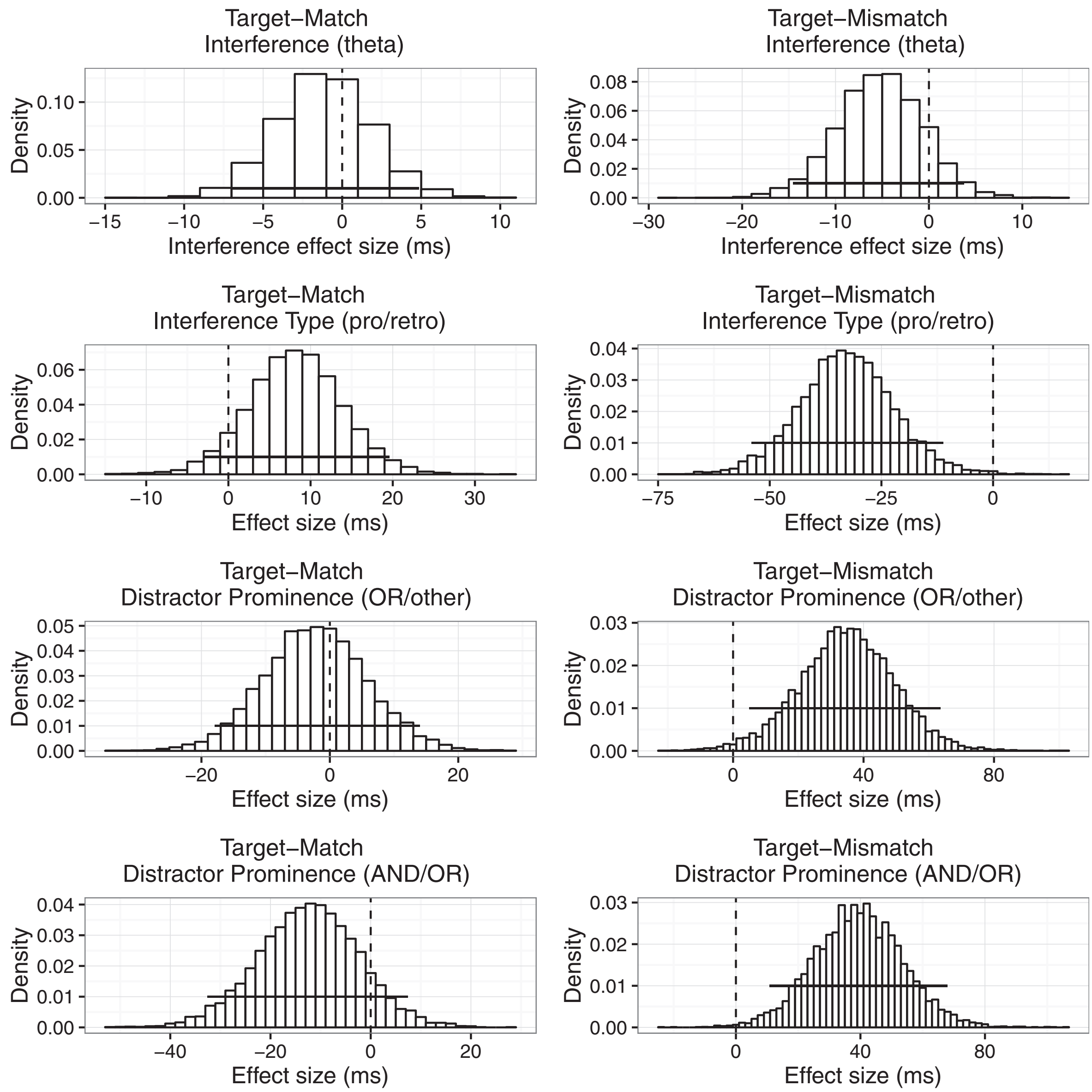


Fig. 4. Summary of the posterior distributions of the interference effect and the effects of interference type (proactive vs. retroactive) and distractor prominence (Other vs. Subject OR Topic, and Subject OR Topic vs. Subject AND Topic) investigated in the meta-regression modeling all dependency types. A positive interference effect means inhibitory interference, whereas a negative effect means facilitatory interference. A positive effect of interference type means that proactive interference leads to a right-shift of the interference effect (i.e., increased inhibition or reduced facilitation) and retroactive interference to a left-shift. A positive effect of the distractor prominence comparisons means that a more prominent distractor leads to a right-shift of the interference effect (i.e., increased inhibition or reduced facilitation). The column on the left shows target-match, and the column on the right shows target-mismatch configurations. The horizontal lines represent 95% credible intervals.

facilitatory interference effect in target-mismatch configurations predicted to attenuate—i.e., a right-shift is predicted. For a computational investigation of the possible effects of distractor prominence, see Engelmann, Jäger, and Vasishth, "The effect of prominence and cue association in retrieval processes: A computational account" (manuscript under review).

One open issue to be addressed in future research is whether the effect of distractor prominence differs between dependency types. We were not able to model distractor prominence in the by-dependency-type analysis because not enough data are available for each level of the prominence factor. Hence, the conclusions above should be treated as preliminary.

Interference type

The by-dependency-type analyses reveal a surprising difference between the dependency types. The type of interference (proactive vs. retroactive interference) affects non-agreement subject-verb dependencies in the opposite way as compared to subject-verb agreement and reflexive-/reciprocal-antecedent dependencies. The meta-analysis shows—albeit rather weak—evidence that in non-agreement subject-verb dependencies (target-match configurations), retroactive interference induces stronger inhibition than proactive interference. This finding is in line with the results of Van Dyke and McElree (2011) who directly compared pro- and retroactive interference configurations within the same experiment. In target-match configurations of subject-verb agreement and reflexive-/reciprocal-antecedent dependencies, in contrast, the meta-analysis reveals that retroactive interference is associated with a left-shift of the interference effect, i.e., a retroactive distractor induces less inhibition or more facilitation than a proactive one. In the case of reflexives, this pattern is not obvious from the qualitative literature review where no indications for an effect

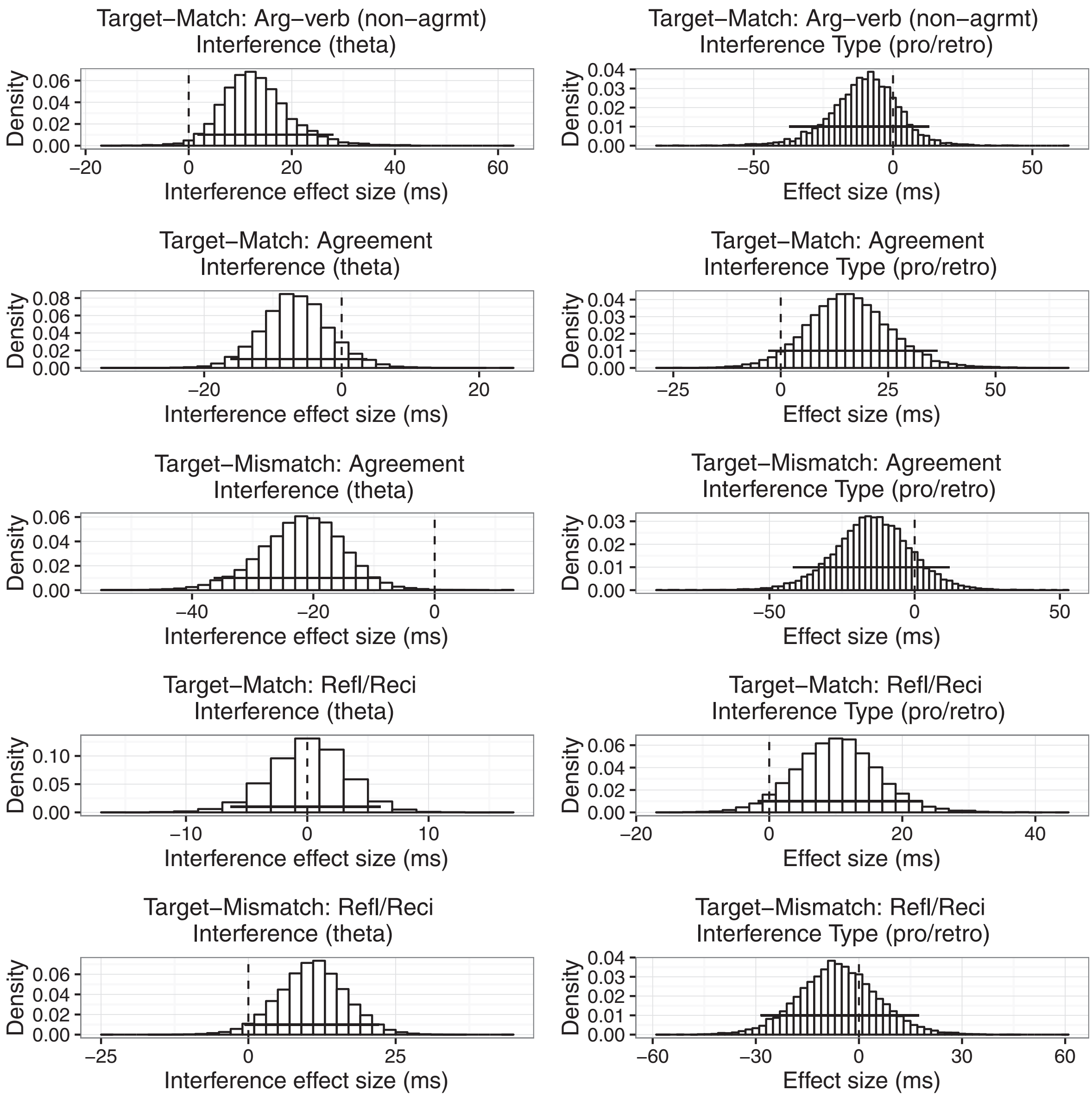


Fig. 5. Summary of the posterior distributions of the interference effect and the effect of interference type (proactive vs. retroactive) investigated in the meta-regressions modeling each dependency type separately. A positive interference effect means inhibitory interference, whereas a negative effect means facilitatory interference. A positive effect of interference type means that proactive interference leads to a right-shift of the interference effect (i.e., increased inhibition or reduced facilitation) and retroactive interference to a left-shift. The horizontal lines represent 95% credible intervals. Target-mismatch configurations are presented only for subject-verb agreement and reflexive/reciprocal dependencies as no data are available for non-agreement subject-verb dependencies. Distractor prominence was not included in this analysis due to sparsity of the data.

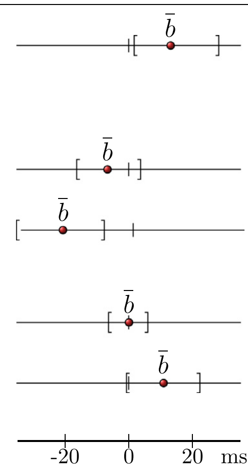
of interference type are found. In the case of subject-verb agreement, this result confirms an observation made in the qualitative literature review: the number of studies observing significant facilitatory effects in target-match configurations of subject-verb agreement (singular verbs) suggests that the facilitation is stronger in a retroactive interference configuration.

In target-mismatch configurations in reflexives/reciprocals and in subject-verb agreement, there is some evidence that a retroactively interfering distractor leads to a right-shift of the interference effect, i.e., more inhibition or less facilitation.

The predictions of the LV05 model with respect to interference type depend on (i) the parameter settings of the model and (ii) the specific experimental materials. The reason for this is that the manipulation of interference type affects mechanisms inducing facilitation as well as mechanisms causing inhibition. On the one hand, the model can predict a left-shift of the interference effect's posterior distribution due to retroactive interference. This is because, everything else being equal, a retroactive interference configuration leads to a higher base level activation of the distractor as the latter has less time to decay than in a proactive interference configuration. This higher activation level of the distractor is predicted to result in a higher proportion of misretrievals of the latter leading to increased facilitation. On the other hand, the model can predict the opposite effect, namely a right-shift of the interference effect's posterior distribution due to retroactive interference. The reason for this is that the target is further away from the retrieval site in most materials testing retroactive interference configurations than in materials examining proactive configurations. This larger distance between target and retrieval site leads to a lower activation level of the target once the retrieval site is reached due to decay. Moreover, in materials with retroactive interference, there is usually more material preceding the retrieval

Table 5
Comparison of the results of the meta-analysis with the predictions of cue-based retrieval as implemented in the LV05 ACT-R model. The Evidence column shows the estimated effect $\bar{b}$ (i.e., the mean of the posterior distribution of the interference effect) together with a 95% credible interval. For the numerical values, see Table 4. The last column summarizes the predictions of the LV05 model and compares it to the results of the meta-analysis. Note that for subject-verb non-agreement dependencies, no data are available for target-mismatch configurations.

Dependency	Target	Evidence	LV05 prediction
Subject-verb (non-agreement)	Match		inhibition ✓
Subject-verb agreement	Match		inhibition ✗
	Mismatch		facilitation ✓
Reflexive-/reciprocal-antecedent	Match		inhibition ✗
	Mismatch		facilitation ✗



site that potentially causes similarity-based interference at the moment of retrieving the target, reducing its activation even more. This lower activation of the target is predicted to increase the inhibitory interference effect. In sum, the specific characteristics of the materials being tested need to be taken into account when deriving predictions of the LV05 model with respect to interference type. Moreover, specific model parameters such as the decay parameter, which determines how fast an item decays in memory, have an important impact on the model's predictions. Hence, computational simulations are needed to derive the ACT-R predictions concerning the effect of interference type.

Possible explanations for the differences between dependency types

We have shown that there are important differences between the dependency types. This is unexpected under the LV05 model, as an implicit assumption in the LV05 model is that the retrieval mechanism needed for the computation of a syntactic dependency is similar for all dependencies.

There are different ways to explain the observed differences between dependency types. The first is that (some of) the differences might arise from methodological issues of various kinds. The second is that the cognitive mechanisms involved in the processing of syntactic dependencies may differ between dependency types.

We first discuss possible methodological issues that may explain the observed differences between dependency types. The particularities of the experimental designs in which the different dependencies were tested might have confounded the data. In particular, the evidence for facilitatory interference in target-match configurations of subject-verb agreement dependencies despite the strong evidence for inhibition in non-agreement subject-verb dependencies may be explained by a confound in the experiments on subject-verb agreement. As mentioned in the qualitative review (see section '*Potential confounds*'), Wagers et al. (2009) pointed out that the pattern observed in subject-verb agreement might be confounded by a spilled-over length/complexity effect of the distractor. In experiments testing singular verbs, the distractor is singular in the distractor-match condition. As a singular distractor is both shorter and morphologically less complex than a plural-marked distractor, it is read faster. This singular-distractor advantage might still continue to affect reading times at the verb, which in many of the reviewed experiments directly follows the distractor. This facilitation in the distractor-match condition might cancel out an inhibitory interference effect. Importantly, all but one of the agreement experiments included in the meta-analysis of target-match configurations tested singular verbs. Hence, to decide whether the facilitatory interference observed in target-match configurations of subject-verb agreement is a valid finding, more experiments are needed in which the above discussed confound is taken care of.

The absence of interference in target-match configurations of reflexives/reciprocals remains unclear. Furthermore, we do not see any confound in the stimuli that might explain the contrasting interference patterns in target-mismatch configurations (facilitation in subject-verb agreement vs. inhibition in reflexives and reciprocals).

It is certainly possible that not all the differences between dependency types reduce to methodological issues. If so, the data presented here may in fact indicate that there are qualitative differences in how the dependency types are processed. This raises the question whether the memory access mechanism per se differs between dependency types as suggested by Dillon (2011), who proposed a syntactically guided serial search mechanism for the memory access of a reflexive's antecedent and a cue-based retrieval mechanism for subject-verb agreement dependencies, or whether the differences between the dependency types can be explained by different properties of the retrieval cues, possibly within the framework of cue-based retrieval. For example, Kush (2013) demonstrated that the usage of *c-command* as a retrieval cue is computationally more complex than the usage of *number* or *gender*. In essence, the problem with relational syntactic features such as *c-command* is that they need to be updated with each incoming word. It might be the case that the human parser therefore more heavily relies on heuristics (e.g., one could approximate *c-command* by the features *subject* and being inside the *local* clause—the *local* feature only needs to be updated whenever the respective clause boundary is crossed; the *subject* feature is constant). A different interference pattern may result from the usage of such heuristic cues. In sum, if it turns out that the differences between the dependency types reduce to differences in the respective retrieval cues, this might be explained within the framework of a content-addressable memory architecture such as ACT-R. If, by contrast, it turns out that the access mechanisms themselves differ between the dependency types in a qualitative manner, this would be incompatible with the way cue-based retrieval theory is currently applied. More theoretical and empirical work is needed to pin down and explain the differences between dependency types.

Possible reasons for the LV05 model's failure to explain the data

As we have shown, the empirical evidence is only partially consistent with the LV05 ACT-R model of cue-based retrieval. However, we want to emphasize that for a more detailed evaluation and discussion of the ACT-R model, computational simulations are needed. In particular, simulations are needed for a quantitative investigation of the interaction between facilitation caused by misretrievals and inhibition due to similarity-based interference. For detailed simulations, see Engelmann, Jäger, and Vasishth, "The effect of prominence and cue association in retrieval processes: A computational account" (manuscript under review).

Generally speaking, there are two ways to explain the model's failure to capture the data. On the one hand, the reason might be that the LV05 model is (at least partially) wrong and therefore not able to capture all facets of linguistic behavior. Given the success of the ACT-R model in other areas of human cognition, it might be premature to completely discard the model. But this meta-analysis clearly shows the need to revise the LV05 ACT-R model of sentence processing. A revised ACT-R model should

(i) explain the differences between the dependency types. The LV05 model currently assumes that the dependency formation works in a similar way for all kinds of dependencies, an assumption is clearly not supported by the data.
(ii) be able to account for inhibitory interference effects in the absence of cue-overload, which was observed in the processing of target-mismatch configurations of reflexives and reciprocals.
(iii) explain the facilitatory (or at least the absence of inhibitory) interference in target-match configurations of subject-verb agreement dependencies as well as the absence of inhibitory interference in target-match configurations of reflexives and reciprocals.

If, however, future research shows that cue-based retrieval is not involved in any dependency resolution process, then the ACT-R model of sentence processing needs to be discarded since cue-based retrieval is one of the core assumptions of the architecture (Anderson & Lebiere, 1998; Anderson et al., 2004).

A second possible reason why the LV05 model fails to capture the empirical evidence is related to the validity of the empirical findings. On the one hand, it might be the case that the observed patterns are driven by confounds in the experimental stimuli (see discussion above and section '*Potential confounds*' of the qualitative literature review). On the other hand, low statistical power of many of the reviewed experiments may have led to misleading estimates of the effects; these are so-called Type S (sign) and M (magnitude) errors. Type S errors refer to situations where the estimated direction of an effect is opposite of the direction of the real effect. Type M errors refer to estimated effect sizes that are much larger than the real effect. For further discussion, see Gelman and Carlin (2014) and section B of the appendix.

Methodological discussion

Discrepancies between the qualitative review and the quantitative analysis

As pointed out above, there are a few important inconsistencies between the conclusions one might be inclined to draw from the qualitative literature review and the quantitative results of the Bayesian meta-analysis. First, and more generally speaking, these inconsistencies show that it is misleading to draw conclusions based on multiple studies using statistical significance as a criterion. This issue is particularly important in a field like psycholinguistics where it is quite common to run experiments with as few as 24 participants. Even though null results based on such small sample sizes are difficult to interpret, and even though this misuse of null results has long been criticized by statisticians,[22] the failure to find a statistically significant effect continues to be widely misinterpreted as demonstrating that there is no effect.[23] Of course, such null results may be very informative and therefore must be published, otherwise publication bias would result. By contrast, in the quantitative meta-analysis every study's estimate is weighted by its uncertainty estimate (the estimated standard error); the meta-analysis is a measurement error model (Gelman et al., 2014). Thus, the Bayesian meta-analysis allows us to pool together all the information contained in any (published) study, i.e., also the ones with non-significant results.

Second, in our analyses in particular, the differences might also be due to the fact that (i) we used first-pass reading time in the quantitative modeling, whereas in the qualitative review we considered any dependent variable reported as significant by the authors of the respective publication, and (ii) in a few cases, the numbers reported in the published paper—which we based the qualitative review on—differed from the means in the raw data that we used for the modeling.

Some limitations of this meta-analysis

As a closing word, we want to emphasize that the results of the Bayesian meta-analysis should be treated with caution. Although the meta-analysis allows us to quantitatively synthesize the existing evidence, it is obviously not able to remedy problems inherent in the data, such as possible confounds in the experimental designs, publication biases of any kind (see section B of the appendix for more details about possible publication bias in the reviewed experiments), etc. Such biases can in principle be quantitatively taken into account in the meta-analysis (for an attempt, see Vasishth, 2015), but this would be a major undertaking beyond the scope of this paper. Moreover, the meta-analysis can only be as good as the numbers that it is based on. As we did not have access to the raw data of many of the included experiments, we had to rely on the condition means and standard errors or standard deviations reported in the respective papers. However, the reported standard deviations in published papers are sometimes computed without taking into account the fact that the data points from one subject (or one item) are not independent, meaning that the reported standard deviation could be an overestimate. Without accessing the raw data, it is not possible to correct for this bias. These overestimates lead to overly conservative estimates of the standard errors.

Lastly, although interference in dependency resolution has attracted considerable attention in the past 15 years and, compared

[22] For example, Neyman (1955), cited in Mayo and Spanos (2006), writes: "[If] the chance of detecting the presence [of discrepancy from the null], . . .is extremely slim, even if [the discrepancy is present]. . ., the failure of the test to reject H0 cannot be reasonably considered as anything like a confirmation of H0. The situation would have been radically different if the power function [corresponding to a discrepancy of interest] were, for example, greater than 0.95."

[23] It may well turn out that the underlying effects are near zero, but this can only be established by conducting replications and high-powered studies.

to other psycholinguistic research topics, the number of published experiments is relatively large, most of the published experiments have low statistical power due to their small sample sizes (see Appendix B for some estimates of power for interference effects in target-mismatch configurations of subject-verb agreement). In such a situation, meta-analysis is a very important tool for evaluating existing evidence. However, low statistical power is known to lead to Type S (sign) and Type M (magnitude) errors (Gelman & Carlin, 2014), which a meta-analysis is not able to remedy. Therefore, we want to emphasize the need to carry out well-designed experiments with higher statistical power, and to attempt to replicate results. Releasing raw data and reproducible code with every publication would also greatly facilitate future work.

Acknowledgements

For partial support of this research, we thank the Volkswagen Foundation through grant 89 953. The statistical analyses of the raw data would not have been possible without the cooperation of Colin Phillips, Brian Dillon, Matt Wagers, Sol Lago, Dave Kush, Matthew Tucker, Ian Cunnings, and Julie Franck; our heartfelt thanks to them. For valuable comments and discussions, we are grateful to Colin Phillips, Patrick Sturt, Ian Cunnings, Brian Dillon, Dan Parker, Dave Kush, Sol Lago, Bruno Nicenboim, and Dario Paape. Thanks go to Richard Morey and Keith O'Rourke for statistical advice and useful pointers to the statistics literature. We also want to thank the audiences of AMLaP 2014, CUNY 2015, and ICCM 2015 for commenting on our work. Finally, we thank the action editor (Klaus Oberauer) and the reviewers (Julie Van Dyke and Marc Brysbaert) for their detailed and helpful comments on previous drafts.

Appendix A. Qualitative literature review: summary tables

Tables A1 and A2.

Appendix B. A note on statistical power and potential publication bias

Low statistical power has been a long-standing problem in psychology (Cohen, 1962), and recent replication failures (Open

Table A1
Reported interference effects for subject-verb dependency comprehension grouped by number of the critical verb and cue-match/mismatch of the target subject.

Publication	Lang.	Method (Measure)	Interf. type	Distractor position	Singular verb				Plural verb			
					Target-Match		Target-Mismatch		Target-Match		Target-Mismatch	
					Effect	AOI	Effect	AOI	Effect	AOI	Effect	AOI
Number agreement												
Acuña-Fariña et al. (2014)	SP	ET (TFT)	retro	PP	−15	crit	–		−32	crit	–	
		(Reg in)			−5%	crit	–		−1%	crit	–	
		(cumRT)			−25	crit	–		−59	crit	–	
		(FPRP)			−5%	crit	–		−5%	crit	–	
		(FFD)			−7	post	–		*n.s.*		–	
*Dillon et al. (2013), Exp1 agrmt	EN	ET (TFT)	retro	obj	*n.s.*		–		–		−118	crit
*Franck et al. (2015), Exp1 compl	FR	SPR	pro	obj	*n.s.*		–		–		–	
*Franck et al. (2015), Exp1 RC	FR	SPR	pro	obj	≈ +100	crit	–		–		–	
Kaan (2002)	NL	ERP (500–700 ms)	retro	obj	*n.s.*		+6.41 μV	crit	*n.s.*		*n.s.*	
*Lago et al. (2015), Exp1	SP	SPR	pro	subj	*n.s.*		–		–		−39	post
*Lago et al. (2015), Exp2	EN	SPR	pro	subj	*n.s.*		–		–		−36	post
*Lago et al. (2015), Exp3a	SP	SPR	pro	subj	−12	post	–		–		−15	post
*Lago et al. (2015), Exp3b	SP	SPR	pro	subj	*n.s.*		–		–		−21	post
Nicol et al. (1997), Exp1	EN	maze	retro	PP	−70	crit	–		*n.s.*		–	
Nicol et al. (1997), Exp2	EN	sent. class.	retro	PP	−124	sent	–		*n.s.*		–	
Nicol et al. (1997), Exp4	EN	sent. class.	retro	PP	−60	sent	–		–		–	
Nicol et al. (1997), Exp5 hg-attchmt	EN	sent. class.	retro	obj[hgRC]	−67	sent	–		–		–	
Nicol et al. (1997), Exp5 lw-attchmt	EN	sent. class.	retro	obj[lwRC]	*n.s.*		–		–		–	
*Pearlmutter et al. (1999), Exp1	EN	SPR	retro	PP	−35	crit	–		–		+19	crit
					−36	post	–		–		−26	post
*Pearlmutter et al. (1999), Exp2	EN	ET (TFT)	retro	PP	−49	post	–		–		−106	crit
											−44	post
		(FPRT)			−36	post	–		–		*n.s.*	
		(FPRP)			*n.s.*		–		–		−15%	crit
*Pearlmutter et al. (1999), Exp3	EN	SPR	retro	PP	−36	crit	–		+24	post	–	
Pearlmutter (2000), Exp1 [*2nd distr.*]	EN	SPR	retro	PP,PP	−23	crit	–		–		–	
Pearlmutter (2000), Exp2 [*1st distr.*]	EN	SPR	retro	PP,PP	–		–		−19	crit	–	
*Tucker et al. (2015)	AR	SPR	retro	obj	−14	post	–		–		−57	crit
*Wagers et al. (2009), Exp2	EN	SPR	pro	subj	*n.s.*		–		–		−58	post
*Wagers et al. (2009), Exp3	EN	SPR	pro	subj	*n.s.*		*n.s.*		*n.s.*		(*facil*)	post
*Wagers et al. (2009), Exp4	EN	SPR	retro	PP	−17	crit	–		–		−32	post
*Wagers et al. (2009), Exp5	EN	SPR	retro	PP	*n.s.*		–		–		−58	post
*Wagers et al. (2009), Exp6	EN	SPR	retro	PP	*n.s.*		–		–		–	
Semantic cues												
*Van Dyke (2007), Exp1 LoSyn	EN	SPR	retro	PP	+54	crit	–		–		–	
*Van Dyke (2007), Exp2 LoSyn	EN	ET (FPRT)	retro	PP	+44	post	–		–		–	
		(RPD)			(+23)	crit	–		–		–	
					+235	post	–		–		–	

Table A1 (*continued*)

Publication	Lang.	Method (Measure)	Interf. type	Distractor position	Singular verb				Plural verb			
					Target-Match		Target-Mismatch		Target-Match		Target-Mismatch	
					Effect	AOI	Effect	AOI	Effect	AOI	Effect	AOI
*Van Dyke (2007), Exp3 LoSyn	EN	ET (RPD)	retro	PP	(−54)	post	–		–		–	
*Van Dyke and McElree (2006)	EN	SPR	pro	mem	+38	crit	–		–		–	
*Van Dyke and McElree (2011), Exp1a pro	EN	SAT (d')	pro	subj	−0.16 (*inhib*)	crit	–		–		–	
Van Dyke and McElree (2011), Exp1a retro	EN	SAT (d')	retro	subj	−0.27 (*inhib*)	crit	–		–		–	
*Van Dyke and McElree (2011), Exp1b pro	EN	ET (TFT)	pro	subj	+20	crit	–		–		–	
*Van Dyke and McElree (2011), Exp1b retro	EN	ET (TFT)	retro	subj	+81	crit	–		–		–	
Van Dyke and McElree (2011), Exp2a pro	EN	SAT (d')	pro	obj	*n.s.*		–		–		–	
Van Dyke and McElree (2011), Exp2a retro	EN	SAT (d')	retro	obj	*n.s.*		–		–		–	
*Van Dyke and McElree (2011), Exp2b pro	EN	ET (TFT)	pro	obj	*n.s.*		–		–		–	
*Van Dyke and McElree (2011), Exp2b retro	EN	ET (TFT)	retro	obj	*n.s.*		–		–		–	
Structural cues (subj.)												
*Van Dyke (2007), Exp1 LoSem	EN	SPR	retro	PP/subj	*n.s.*		–		–		–	
*Van Dyke (2007), Exp2 LoSem	EN	ET (FPRT)	retro	PP/subj	+37	crit	–		–		–	
		(RPD)			+140	crit	–		–		–	
					+395	post	–		–		–	
		(TFT)			+180	crit	–		–		–	
		(FPRP)			+6%	crit	–		–		–	
*Van Dyke (2007), Exp3 LoSem	EN	ET (FPRT)	retro	PP/subj	+20	crit	–		–		–	
		(RPD)			+40	crit	–		–		–	
*Van Dyke and Lewis (2003), Exp4 unambig	EN	SPR	retro	PP/subj	+56	crit	–		–		–	

Note. The experiments are ordered by dependency and cue type (number agreement, semantic cue, structural cue). Those studies marked with an asterisk are included in the meta-analysis. The columns named Effect represent the estimated means of the effect as derived from the numbers reported in the respective publication; the unit is milliseconds in reading studies. Positive values indicate inhibition, negative values indicate facilitation (an exception is the *d'* measure in SAT experiments which indicates facilitation when it is positive). Marginal effects are in parentheses. Effects that are reported as non-significant by the authors are denoted with *n.s.*; '–' means that the respective manipulation was not tested. The experiments are further classified by language (AR = Arabic, EN = English, FR = French, NL = Dutch, SP = Spanish), experimental method (SPR = self-paced reading, ET = eyetracking-while-reading, ERP = event-related potentials, SAT = speed-accuracy tradeoff, sent. class = sentence classification, maze = maze task), interference type (pro = proactive, retro = retroactive), and by syntactic position of the distractor (subj = subject, obj = object, PP = inside a prepositional phrase, mem = three sentence external memory load words). For reading studies, the interest area (AOI) column indicates at which region the effect was observed (crit = critical region containing the verb, post = post-critical region). Whenever an effect is presented as not significant (*n.s.*), this means that it was neither significant at the critical nor at the post-critical region. For eyetracking and SAT experiments, Measure indicates the dependent variable in which the effect was observed (FPRT = first-pass reading time, FFD = first-fixation duration, FPRP = first-pass regression probability, RPD = regression-path duration, TFT = total fixation time, Reg in = proportion of trials with a regression into the region, cumRT = cumulative reading time); for ERP experiments it specifies the time window in which the effect was observed. For those experiments where only a subset of the conditions was considered, the relevant condition labels are specified in the publication column.

Science Collaboration, 2015) have once again brought this issue back into focus. Low power has two major consequences. First, null results tend to be generated. Second, results that do reach significance despite low power are likely to have exaggerated effect sizes due to Type M (magnitude) errors or even go in the wrong direction, so-called Type S (sign) errors (Gelman & Carlin, 2014). Publication bias arises when null results go unpublished due to the tendency to only publish statistically significant results (Rosenthal, 1979), and the tendency to publish significant results that are consistent with theory.

In addition to publication bias, other practices, such as flexibility in data collection and analysis (Simmons, Nelson, & Simonsohn, 2011), lead to so-called reporting bias (Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014). Because these sources of bias usually cannot be retraced once a study is published, it is in general impossible to definitively establish the presence of bias, or to quantify the bias in order to correct for it. In medicine, there have been attempts to take bias into account by eliciting expert opinion (Turner, Spiegelhalter, Smith, & Thompson, 2008), but these methods are extremely time-consuming and difficult to use in practice.

Publication bias can, however, be investigated through several methods. One common graphical approach is to use funnel plots (Egger, Smith, Schneider, & Minder, 1997; Light & Pillemer, 1984). These show the precision (inverse of the square of the standard error) against the observed magnitude of the effect. In the absence of any publication bias, low power studies have a wide, even spread about the true mean (due to Type M and S error), and the higher power studies are progressively more clustered near the true mean. Any gaps or asymmetry in the shape of the resulting funnel plot could be due to publication bias. Note that low precision entails low power, i.e., low power studies are at the lower end of the precision axis and high power studies appear higher up.

The funnel plots in Fig. B1 show some possible indication of asymmetry in the data we have investigated in this paper. In target-match configurations, this asymmetry is present in studies with lower precision (the cluster of studies at the lower end of the funnel): we see a gap in the left-hand side of the base of the funnel plot. It is difficult to determine why this asymmetry exists; we avoid speculation here. In target-mismatch, most of the studies are also clustered in the lower end of the precision range; but we see a fairly symmetric spread around the base of the funnel plot. Thus, the distribution in both target-match and target-mismatch configurations has the characteristic spread associated with low-powered studies.

Table A2
Reported interference effects in reflexive/reciprocal-antecedent comprehension, grouped by cue-match/mismatch of the target antecedent.

Publication	Lang.	Method (Measure)	Cue	Interf. type	Distractor position	Target-match		Target-mismatch	
						Effect	AOI	Effect	AOI
Reflexives (direct object)									
*Jäger et al. (2015), Exp1	CN	ET (FPRT)	anim	retro	subj	*n.s.*		+19	crit
		(RBRT)				*n.s.*		+18	crit
		(FFD)				*n.s.*		+13	crit
		(RPD)				*n.s.*		+10	crit
*Jäger et al. (2015), Exp2 local	CN	ET (FPRT)	anim	pro	mem	+15	crit	–	
		(RBRT)				+19	crit	–	
		(RPD)				+67	crit	–	
		(TFT)				+42	crit	–	
*Felser et al. (2009), Exp2b natives, inacc-mism	EN	ET (RPD)	c-com	pro	subj(topic)	+28	crit	–	
*Badecker and Straub (2002), Exp3	EN	SPR	gend	pro	subj	+42	post	–	
*Badecker and Straub (2002), Exp5	EN	SPR	gend	pro	gen	*n.s.*		–	
*Badecker and Straub (2002), Exp6	EN	SPR	gend	pro	prep obj	*n.s.*		–	
*Cunnings and Felser (2013), Exp1 high WM	EN	ET	gend	pro	subj(topic)	*n.s.*		*n.s.*	
*Cunnings and Felser (2013), Exp1 low WM	EN	ET	gend	pro	subj(topic)	*n.s.*		*n.s.*	
*Cunnings and Felser (2013), Exp2 high WM	EN	ET	gend	retro	subj(topic)	*n.s.*		*n.s.*	
*Cunnings and Felser (2013), Exp2 low WM	EN	ET (FPRT)	gend	retro	subj(topic)	−49	crit	*n.s.*	
		(FFD)				−24	crit	(+23)	crit
*Cunnings and Sturt (2014), Exp1	EN	ET (FPRT)	gend	pro	subj(topic)	*n.s.*		(+39)	post
*Felser et al. (2009), Exp2b natives, no c-com	EN	ET	gend	pro	subj(topic)	*n.s.*		–	
King et al. (2012), adjacent	EN	ET	gend	retro	obj	*n.s.*		*n.s.*	
Nicol (1988)	EN	Primg	gend	pro	subj,obj	*n.s.*		–	
*Patil et al. (2016)	EN	ET (FPRP)	gend	retro	subj	+6.74%	crit	*n.s.*	
		(regr-cont FFD)				(≈ +20)	crit	(≈ −50)	crit
*Sturt (2003), Exp1	EN	ET (RRT)	gend	pro	subj(topic)	−97	post	*n.s.*	
*Sturt (2003), Exp2	EN	ET	gend	retro	obj(topic)	*n.s.*		*n.s.*	
Xiang et al. (2009)	EN	ERP	gend	retro	subj	–		*n.s.*	
*Dillon et al. (2013), Exp1 refl	EN	ET	num	retro	obj	*n.s.*		*n.s.*	
*Dillon et al. (2013), Exp2 himself	EN	ET	num	retro	obj	*n.s.*		*n.s.*	
*Dillon et al. (2013), Exp2 themselves	EN	ET	num	retro	obj	*n.s.*		*n.s.*	
Parker and Phillips (2014), Exp1	EN	ET (TFT)	num/gend	pro	subj	*n.s.*		≈ −250	crit
Parker and Phillips (2014), Exp2	EN	ET (TFT)	num/anim	pro	subj	*n.s.*		≈ −170	crit
Parker and Phillips (2014), Exp3	EN	ET (TFT)	gend/ anim	pro	subj	*n.s.*		≈ −250	crit
Prepositional reflexives									
Clackson et al. (2011), Exp2 adults	EN	VW	gend	pro	subj(topic)	*n.s.*		–	
Clackson and Heyer (2014)	EN	VW (target ident.)	gend	pro	subj(topic)	*inhib*	200–600 ms	–	
King et al. (2012), non-adjacent	EN	ET (FPRT)	gend	retro	prep obj	*n.s.*		≈ −95	crit
Possessive reflexives									
*Chen et al. (2012), local	CN	SPR	anim	retro	subj	*n.s.*		–	
Reciprocals									
*Badecker and Straub (2002), Exp4	EN	SPR	num	pro	subj	+48	post	–	
*Kush and Phillips (2014)	HI	SPR	num	retro	prep obj	*n.s.*		(+30)	post

Note. The experiments are ordered by dependency type (direct object reflexives, prepositional reflexives, possessive reflexives, reciprocals) and investigated cue (anim = animacy, c-com = c-command, gend = gender, num = number). Those studies marked with an asterisk are included in the meta-analysis. The columns named Effect represent the estimated means of the effect as derived from the numbers reported in the respective publication; the unit is milliseconds in reading studies. Positive values indicate inhibition, negative values indicate facilitation. Marginal effects are in parentheses. Effects reported as non-significant by the authors are denoted with *n.s.*; '–' means that the respective manipulation was not tested. The experiments are further classified by language (CN = Mandarin Chinese, EN = English, HI = Hindi), experimental method (SPR = self-paced reading, ET = eyetracking-while-reading, VW = visual world eyetracking, ERP = event-related potentials, Primg = cross-modal priming), interference type (pro = proactive, retro = retroactive), and by syntactic position of the distractor (subj = subject, obj = object, prep obj = prepositional object, gen = genitive attribute, mem = three sentence external memory load words, topic = discourse topic). For reading studies, the interest area (AOI) column indicates at which region the effect was observed (crit = critical region containing the reflexive/reciprocal, post = post-critical region). For visual world studies, the interest area refers to the time window in which the effect was observed. Whenever an effect is presented as not significant (*n.s.*), this means that it was neither significant at the critical nor at the post-critical region. For eyetracking experiments, Measure indicates the dependent variable in which the effect was observed (FPRT = first-pass reading time, FFD = first-fixation duration, regr-cont FFD = regression-contingent first-fixation duration, FPRP = first-pass regression probability, RBRT = right-bounded reading time, RPD = regression-path duration, RRT = re-reading time, TFT = total fixation time, target ident. = target identification). For those experiments where only a subset of the conditions was considered, the relevant condition labels are specified in the Publication column.

What could be done differently when investigating interference and other phenomena in sentence comprehension? Fixing sample size by calculating power functions before carrying out an experiment is an important first step that has generally been neglected in psycholinguistics. Power functions can be estimated by looking at previously published studies or predictions from computational/mathematical models. As an illustration, consider Experiments 2–5 of Wagers et al. (2009), the agreement conditions in Experiment 1 of Dillon et al. (2013), and the four experiments of Lago et al. (2015). All these experiments were carried out by the same research group, and all these studies investigate ungrammatical (target-mismatch) configurations of subject-verb agreement dependencies. These studies can therefore be considered to have a degree of homogeneity that cannot be expected when combining

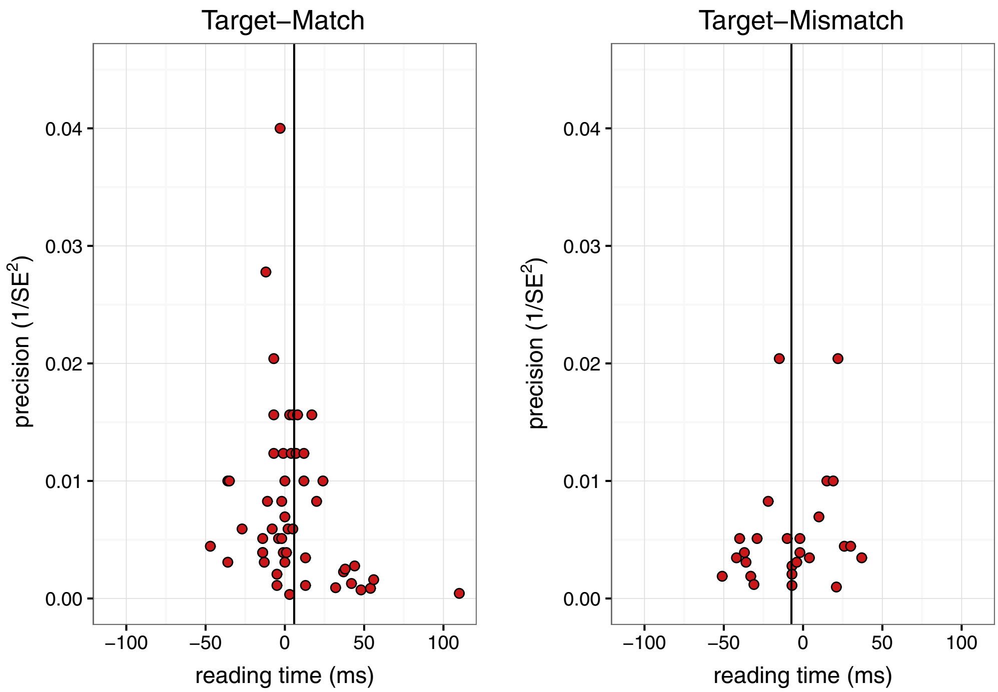


Fig. B1. Funnel plots for target-match and target-mismatch data. The points represent the estimated interference effect from the studies included in the meta-analysis. The vertical line in each plot represents the grand mean of all the data points.

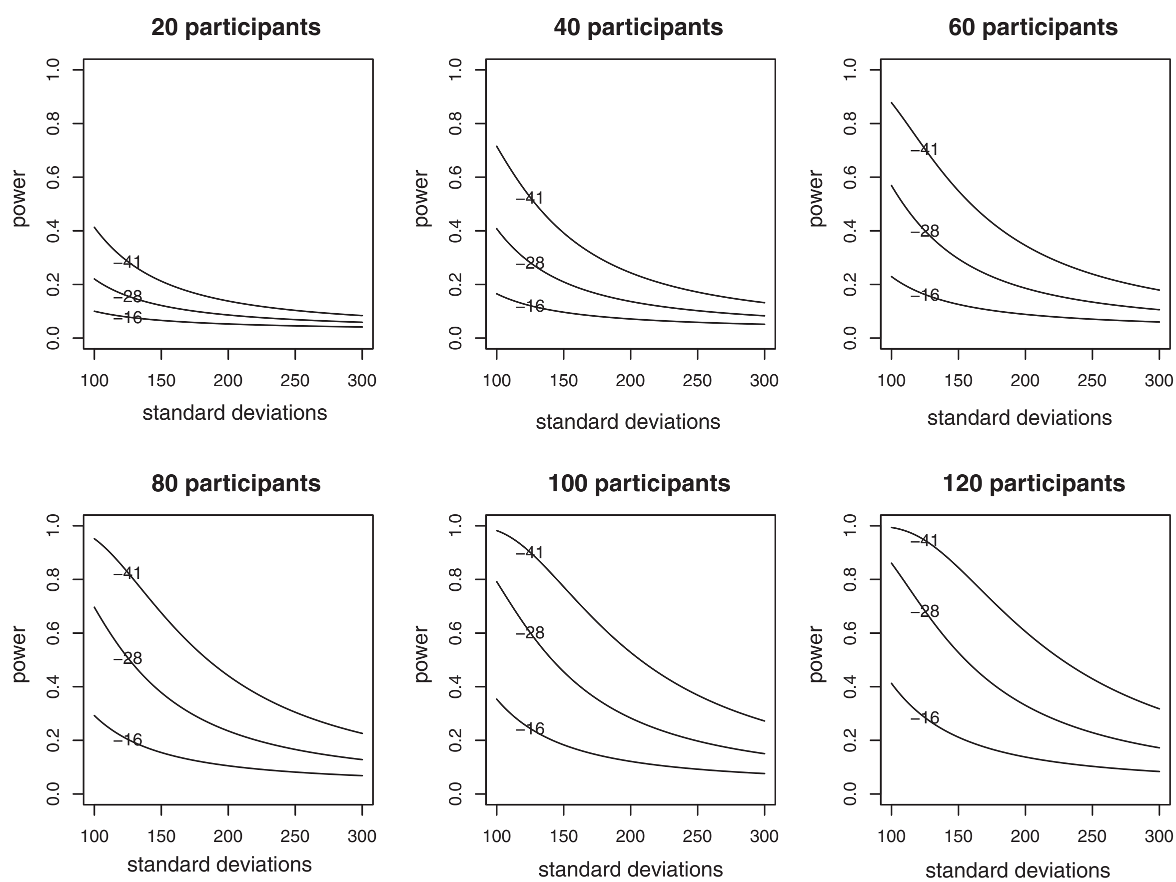


Fig. B2. Power functions for interference in subject-verb agreement dependencies (target-mismatch) based on estimates from published data. This analysis assumes a sample size of 20–120 participants, and standard deviations with the range seen in the studies considered here, and three plausible effect sizes of −16 ms, −28 ms, or −41 ms (based on the meta-analysis of the experiments investigated here).

results from different labs. The homogeneity has the positive consequence that between-study variance is expected to be relatively low. A random-effects meta-analysis of these experiments using Stan, Version 2.14.1 (Stan Development Team, 2016) yields a posterior distribution of the effect with mean −28 ms and 95% credible interval [−41, −16] ms. We can take these estimates as plausible values for the interference effect in target-mismatch configurations of subject-verb agreement in a future study. Taking 100–300 ms as a plausible range of standard deviations for reading studies,[24] we can compute an approximate power function for different sample sizes; see Fig. B2. From the figure, it is clear that for relatively small sample sizes, the power estimates given the range of plausible effect sizes are quite low. Sample sizes such as 100 or 120 participants may be more appropriate in this particular case, given these particular data. Thus, for future work, we suggest that experiments be appropriately powered when investigating this and also other issues, so that Type M and S errors are avoided and accurate estimates can be recorded.

References

Acuña-Fariña, J. C., Meseguer, E., & Carreiras, M. (2014). Gender and number agreement in comprehension in Spanish. *Lingua, 143*, 108–128.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc..

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 748–769.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine, 14*(4), 395–411.

Boston, M., Hale, J., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes, 26*(3), 301–349.

Chen, Z., Jäger, L. A., & Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. In B. Stolterfoht & S. Featherston (Eds.), *Empirical approaches to linguistic theory: Studies of meaning and structure* (pp. 43–62). Berlin, Germany: Mouton de Gruyter.

Chen, D.-G. D., & Peace, K. E. (2013). *Applied meta-analysis with R*. Boca Raton, FL: CRC Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, The Netherlands: Foris.

Clackson, K., Felser, C., & Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: Evidence from eye-movements during listening. *Journal of Memory and Language, 65*, 128–144.

Clackson, K., & Heyer, V. (2014). Reflexive anaphor resolution in spoken language comprehension: Structural constraints and beyond. *Frontiers in Psychology, 5*(904).

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145.

Cunnings, I., & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes, 28*(1–2), 188–219.

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language, 75*, 117–139.

Dillon, B. W. (2011). *Structured access in sentence comprehension*. Ph.D. thesis, University of Maryland, College Park, MD.

Dillon, B. W., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language, 69*, 85–103.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634.

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

Franck, J., Colonna, S., & Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in Psychology, 6*(349).

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651.

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

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Hartsuiker, R. J., Antón-Méndez, I., & van Zee, M. (2001). Object attraction in subject-verb agreement construction. *Journal of Memory and Language, 45*, 546–572.

Higgins, J., & Green, S. (2008). *Cochrane handbook for systematics reviews of interventions*. New York, NY: Wiley-Blackwell.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*(5), 235–241.

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

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: An ERP study. *Journal of Psycholinguistic Research, 31*(2), 165–193.

King, J., Andrews, C., & Wagers, M. (2012). Do reflexives always find a grammatical antecedent for themselves? In *25th Annual CUNY Conference on Human Sentence Processing*. New York, NY: The CUNY Graduate Center, p. 67.

Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing*. Ph.D. thesis. University of Maryland, College Park, MD.

Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology, 5*(1252).

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language, 82*, 133–149.

Lau, E. F., Rozanova, K., & Phillips, C. (2007). Syntactic prediction and lexical surface frequency effects in sentence processing. *University of Maryland Working Papers in Linguistics, 16*, 163–200.

Lehtonen, M., Niska, H., Wande, E., Niemi, J., & Laine, M. (2006). Recognition of inflected words in a morphologically limited language: Frequency effects in monolinguals and bilinguals. *Journal of Psycholinguistic Research, 35*(2), 121–146.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*(3), 375–419.

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*(10), 447–454.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Logačev, P., & Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science, 40*(2), 266–298.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57*(2), 323–357.

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language, 32*, 536–571.

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

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.). *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 155–200). San Diego, CA: Elsevier.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language, 48*, 67–91.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103–123.

New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language, 51*(4), 568–585.

Neyman, J. (1955). The problem of inductive inference. *Communications on Pure and Applied Mathematics, 8*(1), 13–45.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory differences in unbounded dependencies. *Frontiers in Psychology, 7*(280).

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas – Part II. *Language and Linguistics Compass, 10*(11), 591–613.

Nicol, J. (1988). *Coreference processing during sentence comprehension*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Nicol, J., Forster, K. I., & Veres, C. (1997). Subject-verb agreement processes in comprehension. *Journal of Memory and Language, 36*, 569–587.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language, 55*, 601–626.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Parker, D., & Phillips, C. (2014). Selective priority for structure in memory retrieval. In *27th Annual CUNY Conference on Human Sentence Processing*. Columbus, OH: Ohio State University, p. 100.

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition, 157*, 321–339.

Patil, U., Hanne, S., Burchert, F., Bleser, R. D., & Vasishth, S. (2016). A computational evaluation of sentence comprehension deficits in aphasia. *Cognitive Science, 40*, 5–50.

[24] This approximate range of standard deviations was observed in the studies investigated here.

Patil, U., Vasishth, S., & Lewis, R. L. (2016). Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology, 7*(329).

Pearlmutter, N. J. (2000). Linear versus hierarchical agreement feature processing in comprehension. *Journal of Psycholinguistic Research, 29*(1), 89–98.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language, 41*, 427–456.

Phillips, C., Wagers, M., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner (Ed.). *Experiments at the interfaces* (Vol. 37, pp. 147–180). Bingley, UK: Emerald Group Publishing Limited.

Plummer, M. (2012). JAGS version 3.3.0 manual. [Computer software manual]. Retrieved from <http://sourceforge.net/projects/mcmc-jags/files/>.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye movements. *Cognition, 89*(1), B1–B13.

Severens, E., Jansma, B. M., & Hartsuiker, R. J. (2008). Morphophonological influences on the comprehension of subject-verb agreement: An ERP study. *Brain Research, 1228*, 135–144.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Stan Development Team (2013). *Stan: A C++ library for probability and sampling, version 2.1*. Retrieved from <http://mc-stan.org/>.

Stan Development Team (2016). Stan modeling language user's guide and reference manual, version 2.31.1 [Computer software manual]. Retrieved from <http://mc-stan.org/>.

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

Sutton, A. J., Welton, N. J., Cooper, N., Abrams, K. R., & Ades, A. (2012). *Evidence synthesis for decision making in healthcare*. New York, NY: John Wiley & Sons.

Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology, 6*(347).

Turner, R., Spiegelhalter, D., Smith, G., & Thompson, S. (2008). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*(1), 21–47.

Van Dyke, J. (2002). *Parsing as working memory retrieval: Interference, decay, and priming effects in long distance attachment*. Ph.D. thesis. University of Pittsburgh, PA.

Van Dyke, J. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*(2), 407–430.

Van Dyke, J., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language, 49*, 285–316.

Van Dyke, J., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language, 55*(2), 157–166.

Van Dyke, J., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language, 65*(3), 247–263.

Van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language, 45*(2), 225–258.

Vasishth, S. (2015). *A meta-analysis of relative clause processing in Mandarin Chinese using bias modeling*. M.Sc. thesis. School of Mathematics and Statistics, University of Sheffield, Sheffield, UK.

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

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS One, 8*(10), e77006.

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

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*, 125–134.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S-PLUS*. New York, NY: Springer.

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language, 61*, 206–237.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory, 104*(4), 442–452.

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