2 Dependencies in Sentence Comprehension

2.1 Memory Processes in Sentence Comprehension

Cognitive psychology has a rich history of investigating human memory processes. A typical experiment in cognitive psychology might involve getting participants to study pairs of words in succession and then asking them to recall the second word given the first, or to recall the first word given the second. Using such experimental paradigms, psychologists have developed many theories about how human memory works. For sentence processing, especially interesting are theories explaining why we forget material that we have previously seen.

A dominant explanation for forgetting is interference: the association between multiple items in memory leads to competition among them and the subsequent inability to retrieve the correct item. Anderson (1974) suggested that interference is affected by the total number of associated links in memory; he refers to this number as the fan. The larger the fan, the greater the interference. Following Dillon (2011), we will refer to the processing difficulty arising from the fan effect as *inhibitory interference*. Figure 2.1 shows a schematic illustration of the fan effect. Suppose an experiment is carried out where a participant is shown a grey circle, triangle, and square, and the participant's task is to identify the grey square. The participant initiates a search, looking for an object that is grey and a square; these are the cues used for a search. Because there are three items that match the cue grey, the fan is three. Here, identifying the target object (the grey square) will be slower compared to a case where the circle and triangle are not grey.

Apart from interference arising from the fan effect, many other interesting generalizations have emerged from the study of memory in psychology. Some that seem to be important for sentence processing are the following:

(i) Recency and primacy effects (Gibson et al., 1996; Nairne, 1988)
(ii) Pro- and retroactive interference (Keppel and Underwood, 1962; Lewis, 1996; Watkins and Watkins, 1975)
(iii) Misretrieval of items from memory (Patson and Husband, 2016)

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

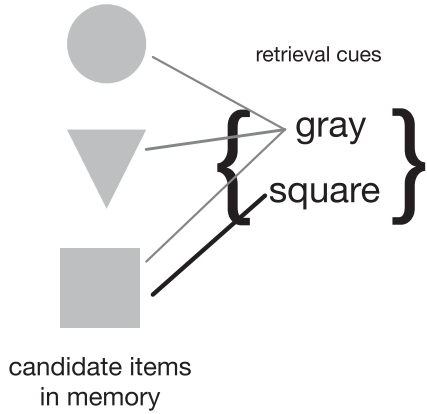

Figure 2.1 A schematic illustration of the fan effect. Searching for an object that is grey and a square (the target item) is more difficult when competing items have one or more features matching cues used for identifying the target item.

(iv) Reactivation of the memorial representation due to repeated accesses from memory (Vasishth and Lewis, 2006)
(v) Richer encoding of an item in memory leading to easier access, due to increased prominence (Hofmeister, 2007, 2011; Hofmeister and Vasishth, 2014)
(vi) Shared features between multiple items in memory, degrading memory representations (Nairne, 1990; Oberauer and Kliegl, 2006; Vasishth et al., 2017a).

Despite this apparently rich connection between sentence processing and the findings of cognitive psychology, it is reasonable to question whether these generalizations from psychology have anything to do with constraints on sentence processing. A priori, the answer could well be: not much. Unlike memorizing unrelated items in a psychology experiment, the words that appear in a sentence occur in a particular context: syntax, semantics, pragmatics, discourse, gesture, prosody, and possibly also facial expressions together impose structure and add context in a way that cannot be compared to simple list memorization and recall. For example, consider a task where a participant is asked to memorize word pairs like

(6) reporter-hired, editor-admitted, article-write

The word associations that the participant would build up in reading these words without any surrounding context will be quite different from the situation when one reads them in a full sentence:

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

(7) The editor who hired the reporter to write the article admitted his mistake

What is similar between word pair memorization and the above sentence is that associations between the same sets of words need to be built. But the differences from word pair memorization stand out: the nature of the associations between the words in the sentence is much richer and more constrained compared to the association in the word pair context. The word associations within a sentence are clearly grounded in a lifetime of experience with the real world and with the grammar of the language in question.

Seen in this way, a priori it seems unlikely that generalizations about memory derived from having participants memorize random lists of words could apply to a structured information processing task such as sentence comprehension. Nevertheless, psycholinguists have investigated the possibility that these generalizations about memory processes also come into play in sentence parsing. Under this view, some, but not all, aspects of the memory system are assumed to affect syntactic structure building and interpretation.

In the next section, we survey the literature on how constraints on memory might play a role in the formation of dependencies between words in a sentence context.

2.2 Dependency Completion in Sentence Processing

Who did what is a central component of comprehending the meaning of a sentence. We will refer to the process of connecting co-dependents for interpretation as dependency completion. This can be seen as a word-association process not unlike the one studied in memory research in psychology, but with the crucial difference that the associations lead to the construction of structured representations.

Completing a dependency crucially involves retrieving a co-dependent element that resides in memory in a heightened state of activation, usually because it has been read or heard in the recent past. This retrieval process is widely assumed to be driven by a cue-based, content-addressable search (McElree et al., 2003). For example, the verb *slept* would generally require an animate subject; one retrieval cue here would therefore be animacy. Another example is number marking on the auxiliary verb; the sentence *The key is on the table* has a singular-marked auxiliary verb which, due to the subject-verb number agreement constraint in English, needs a singular-marked subject.

In such retrieval situations, if more than one noun is present that has a feature that matches the retrieval cue, retrieving the correct noun has been argued to become more difficult. Here, we will refer to the correct target for retrieval as the target noun, and the interfering noun or nouns as the distractor(s).

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

As discussed by Lewis (1996), there are in principle two possible serial order configurations for the target and the distractor(s): the distractor noun can intervene between the target and the verb, or the distractor can precede the target noun. Following the terminology from memory research in cognitive psychology, Lewis (1996) refers to these configurations as proactive and retroactive interference, respectively. As an example, consider the eyetracking study by Van Dyke and McElree (2011). As shown in Example 8, the critical region is the verb *compromised*. Assuming that this verb takes an animate noun as subject, at the verb the animate subject *attorney* must be retrieved. In (8b), the animate distractor noun *witness* appears before the target noun, leading to a proactive interference configuration. The baseline condition here is (8a), which has an inanimate distractor noun *motion*. Retroactive interference arises in (8d) because the distractor *witness* appears between the target noun and the verb; the baseline condition here is (8c).

(8) a. Proactive interference: Low interference
The judge / who had declared that / the **motion** / was inappropriate / realized that the **attorney** / in the case / **compromised** …

b. Proactive interference: High interference
The judge / who had declared that / the **witness** / was inappropriate / realized that the **attorney** / in the case / **compromised** …

c. Retroactive interference: Low interference
The **attorney** / who the judge realized / had declared that / the **motion** / was inappropriate / **compromised** …

d. Retroactive interference: High interference
The **attorney** / who the judge realized / had declared that / the **witness** / was inappropriate / **compromised** …

In the above example, interference is argued to lead to slowdowns at the verb. Pro- and retroactive configurations can also lead to facilitatory effects, if there is a partial match with a proper subset of the retrieval cues triggered at a verb. An example is the study by Wagers et al. (2009). The authors investigated ungrammatical sentences that can lead to illusions of grammaticality due to a partial feature match between plural number marking on the distractor noun and the verb's plural feature. The distractor *musicians* (in the proactive interference condition 9b) and *cells* (in the retroactive condition 9d) can lead to an illusion of grammaticality, leading to faster reading times at the auxiliary (relative to the respective baseline conditions).

(9) a. Proactive interference, distractor mismatch
*The musician who the reviewer praise so highly will …

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

b. Proactive interference, distractor match
*The musicians who the reviewer praise so highly will …
c. Retroactive interference, distractor mismatch
*The key to the cell (unsurprisingly) were rusty from many years of disuse
d. Retroactive interference, distractor match
*The key to the cells (unsurprisingly) were rusty from many years of disuse

Apart from the work mentioned above, pro- and retroactive interference configurations have not been systematically studied in sentence processing; this is an important gap in the literature on interference.

We turn next to a typology of linguistic dependencies that have been investigated in the reading literature. We limit the discussion to work on reading (self-paced reading and eyetracking) because the mapping between reading time and the predictions of the models under consideration in this book is relatively straightforward to define.

Primarily because there is empirical data available from reading studies on these dependency types, we will focus on three basic classes of dependency:

(i) **Subject-verb non-agreement dependencies**. These dependencies involve the grammatical subject of a verb. A simple example would be *The senator read the report.* Here, *senator* is subject of the verb *read*. The reason these are referred to as "non-agreement" dependencies is to distinguish them from subject-verb dependencies in which the verb has an overt number marking that must match that of the subject (at least in subject-agreement languages).
(ii) **Subject-verb number agreement**. As mentioned above, here, the verb carries number marking that must (in the languages considered here, usually English) agree with that of the subject. A classic example is *The keys to the cabinets are on the table*, where the key dependency is that between *keys* and *are*. These subject-verb number agreement constructions are treated separately from the subject-verb non-agreement dependencies because they behave very differently, and the empirical data pose some interesting problems for sentence processing models (especially the ones presented in this book).
(iii) **Reflexives and reciprocals**. These dependencies arise between antecedents and reflexives like English *herself* or Mandarin *ziji*; and reciprocals like English *each other*. These dependencies are special because they involve the binding theory, a central construct in linguistic theory; empirically, what's interesting is that they seem to show distinct patterns of dependency completion time than the other types mentioned above.

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

2.3 Subject-Verb Non-Agreement Dependencies

Julie Van Dyke has carried out a comprehensive set of experiments which suggest that inhibitory interference effects arise when a grammatical subject needs to be connected to a verb and one or more other nouns in memory share certain features with the subject noun.

As an example, consider the self-paced reading study carried out by Van Dyke and McElree (2006). They showed participants sentences like (10). Before participants saw the target sentence, they were either asked to memorize three words, here *table*, *sink*, *truck* (memory-load condition) or not asked to memorize any words (no-memory-load condition). Every sentence was followed by a question like *Did the guy live by the sea?*

(10) a. No-interference condition
It was the boat that the guy who lived by the sea sailed in two sunny days.

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

In the memory-load condition, participants showed longer reading times at the word *fixed* vs. *sailed*; by comparison, in the no-load condition, no difference was seen between the two words. Van Dyke and McElree's conclusion was that this effect was due to increased similarity-based interference at the verb *fixed*: the reader had to retrieve the subject *boat*, but also had three interfering words in memory (*table*, *sink*, *truck*) that could potentially be subjects of the verb *fixed*. These interfering words cause slowdowns in the dependency-completion at the verb. As the authors put it (p. 163): "Reading times for … the locus of the interference manipulation … provided the critical test of our retrieval interference hypothesis. We found clear support for retrieval interference from the significant effect of interference and the significant interaction in this region, which revealed that the interference effect was linked to the difference between the two sentence types in the Load conditions."

One difficulty here is that the claim above is based on a statistically non-significant result (the interaction has min $F'(1,68) = 1.44$, $p = 0.23$; 56 participants), and the interaction between load and interference also failed to show an effect in a subsequent attempt by Van Dyke et al. (2014) (estimate −10 ms, $SE = 15$; 65 participants). In the Van Dyke et al. (2014) paper, there isn't enough information to determine whether the direction of the effect is identical to that of the original study. In a recent larger-sample replication attempt involving English (Mertzen et al., 2020a), we failed to find an interaction between load and interference in total reading time, although we do see some evidence for the interaction in first-pass reading time. Mertzen et al. (2020a)

also carried out eyetracking experiments in parallel on German and Russian, but these languages didn't show any evidence of the expected interaction in any eyetracking dependent measure.

In subsequent work, Van Dyke (2007) conducted eyetracking reading studies in which sentences like (11, 12) were shown. The labels on each sentence type are explained below.

(11) a. LoSyn, LoSem
The worker was surprised that the resident who was living near the dangerous warehouse was complaining about the investigation.

b. HiSyn, HiSem
The worker was surprised that the resident who said that the neighbour was dangerous was complaining about the investigation.

(12) a. HiSyn, LoSem
The worker was surprised that the resident who said that the warehouse was dangerous was complaining about the investigation.

b. LoSyn, HiSem
The worker was surprised that the resident who was living near the dangerous neighbour was complaining about the investigation.

This experiment had a 2 × 2 factorial design which varied whether a noun (*warehouse/neighbour*) that appears between a subject-verb dependency (*resident–was complaining*) was a grammatical subject (High Syntactic interference) or not (Low Syntactic interference), and was animate (High Semantic interference) or not (Low Semantic interference). The research question was the following: when a subject-verb dependency is to be completed at the verb phrase *was complaining*, can a distractor noun (such as *neighbour*) that overlaps in syntactic and semantic features with the grammatical subject (*resident*) cause greater difficulty in completing the dependency at the verb? Van Dyke found that syntactic interference effects occurred earlier than semantic interference effects: when the distractor noun was in subject position inside the relative clause (compared to non-subject position), an interference effect showed up earlier compared to the case where the distractor noun was animate. This suggests that syntactic cues may have priority or may be weighted more heavily than semantic cues.

As mentioned above, Van Dyke and McElree (2011) also investigated interference in proactive and retroactive configurations (see Example 8 above), and argued that retroactive interference effects are stronger than proactive interference effects.

All the experiments by Van Dyke and colleagues investigated sentences with a subject-verb dependency, where the retrieval cues were either syntactic or semantic in nature: in other words, the target noun was either a grammatical

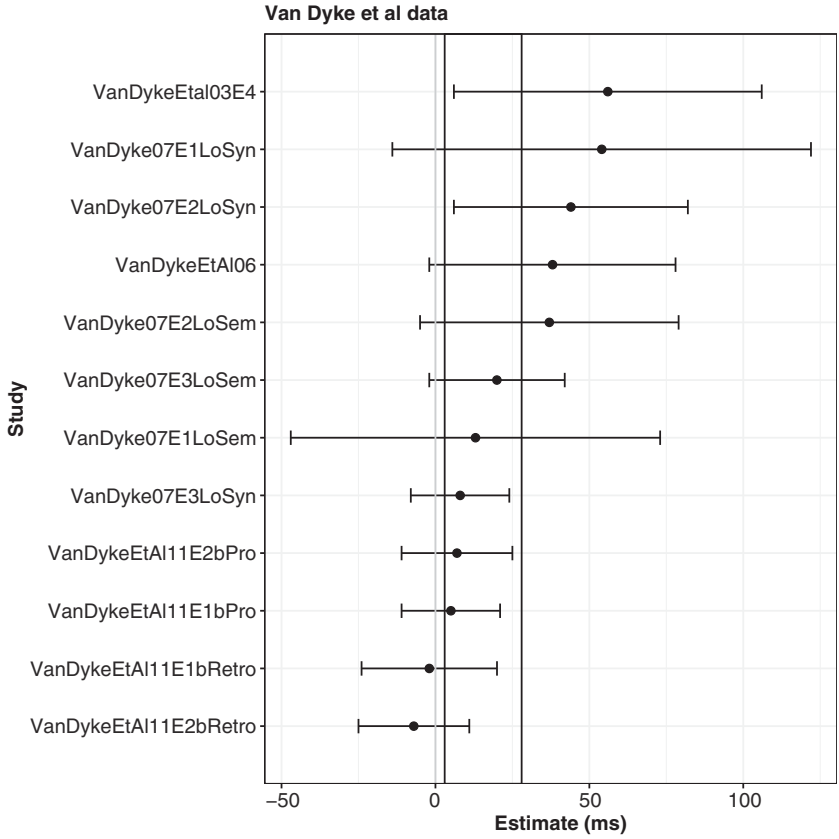

Figure 2.2 Inhibitory interference effects (sorted in increasing order by magnitude) in reading studies by Van Dyke and colleagues. The grey vertical lines show the 95% credible interval for the meta-analysis estimate of the effect.

subject or object, or animate or inanimate. Jäger et al. (2017) assembled the estimates and standard errors for all the studies carried out by Van Dyke and colleagues. As shown in Figure 2.2, these studies tend to show a consistent pattern: with some exceptions, when a distractor noun is present that has features matching the retrieval cues of the verb, an increase in processing time (reading time) is observed. We can summarize these results by computing the posterior distribution of the effect, using a random effects meta-analysis; see Jäger et al. (2017) for details. The meta-analysis shows that the presence of a distractor increases reading time at the verb by 13 ms, with a 95% credible interval of [3,28] ms. Note, however, that a recent pair of eyetracking

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

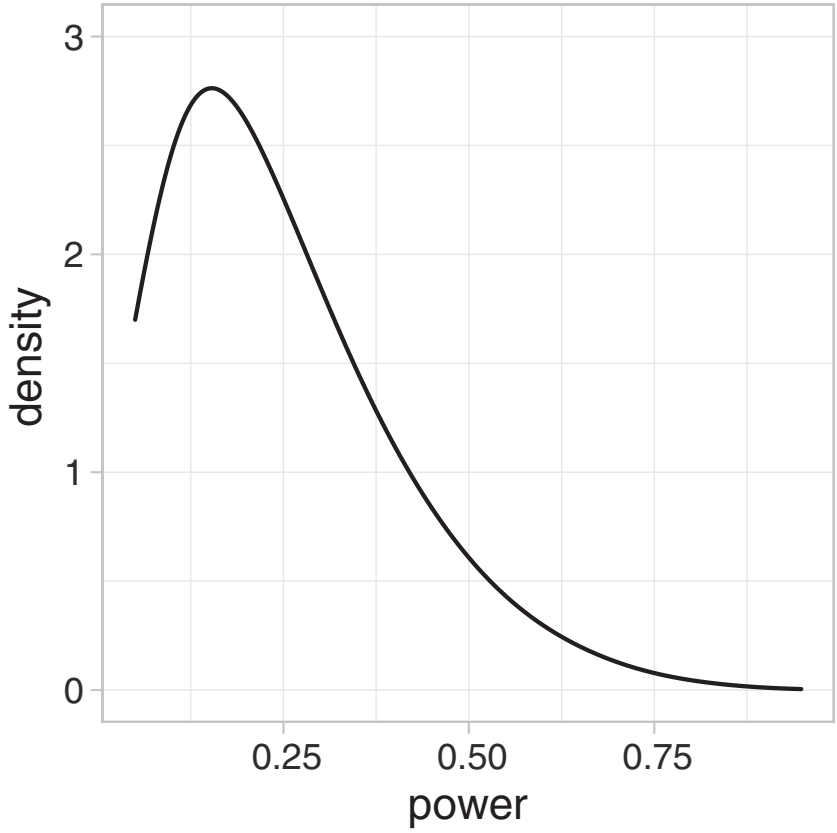

Figure 2.3 Distribution of power (paired, two-sided *t*-test) assuming that the effect has normal distribution with mean 13 and standard deviation 6, the standard deviation ranges from 75 to 100 ms, and the subject sample size is 60.

experiments by Cunnings and Sturt (2018) investigating fan effects found no evidence for inhibitory interference.

Such a failure to find interference effects is no surprise; if the true effect really is in the range [3,28] ms, as Van Dyke's work suggests, then, assuming a standard deviation of 75 ms for reading times (eyetracking or self-paced reading) and a sample size of 60 participants, power is in the range from 6% to 81% (see Figure 2.3 for the power distribution, assuming that standard deviation ranges from 75 to 100 ms and subject sample size is 60). Given that sample sizes are often much lower than 60 participants, power is probably much lower. For example, Cunnings and Sturt had 48 participants; such a sample size would result in power in the range from 6% to 72% for a standard deviation of 75.

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

Thus, with such small sample sizes, an absence of an interference effect is not possible to interpret.

A widely accepted explanation for the inhibitory interference effects is that the retrieval cue cannot uniquely identify the target noun, and this leads to increased processing difficulty due to spreading activation; this is the so-called fan effect (Anderson et al., 2004).

Interestingly, in certain situations, subject-verb dependency configurations can also show facilitatory interference. One plausible explanation for this is a so-called race process (Raab, 1962) triggered by a partial feature match: a subset of the retrieval cues triggered at the verb match with a distractor noun and another subset of cues match with the target, leading to a race process that results in an occasional misretrieval of the distractor (Logačev and Vasishth, 2015; Nicenboim and Vasishth, 2018). The race process is discussed further in Section 3.2.1.

For example, evidence for such a facilitatory interference effect in grammatical subject-verb dependencies comes from Cunnings and Sturt (2018). They conducted two eyetracking (reading) studies in which they manipulated the plausibility of the correct dependent of the verb and the plausibility of the distractor noun. They showed that when the correct dependent is implausible, the distractor's plausibility influences reading time at the verb is faster when the distractor is a plausible subject of the verb. Faster total reading times are observed at the verb *shattered* in (13a) compared to (13b). In their experiment 1, the facilitation effect at the verb was estimated to be -22 ms, $[-4, -42]$, and in experiment 2, it was -19 ms, $[1, -40]$.

(13) a. What Sue remembered was the letter that the butler with the cup accidently shattered today in the dining room.

b. What Sue remembered was the letter that the butler with the tie accidently shattered today in the dining room.

One explanation for this facilitation is in terms of a lognormal race (although this is not how Cunnings and Sturt explain it): the verb *shattered* searches for a subject noun with the property "can be shattered", and in some trials ends up incorrectly retrieving the noun *cup* as the subject; the correct subject is *letter*. Thus, the observed facilitation could be explained by assuming occasional misretrievals of the distractor due to a partial feature match. The process of partial matching leading to occasional misretrievals is graphically summarized in Figure 2.4.

Subject-verb dependencies have also been investigated in the context of number agreement. Here, the retrieval cue of interest is number marking: at least in English, the subject must agree in number with the verb. Dependencies involving number agreement exhibit some interesting peculiarities, as we discuss next.

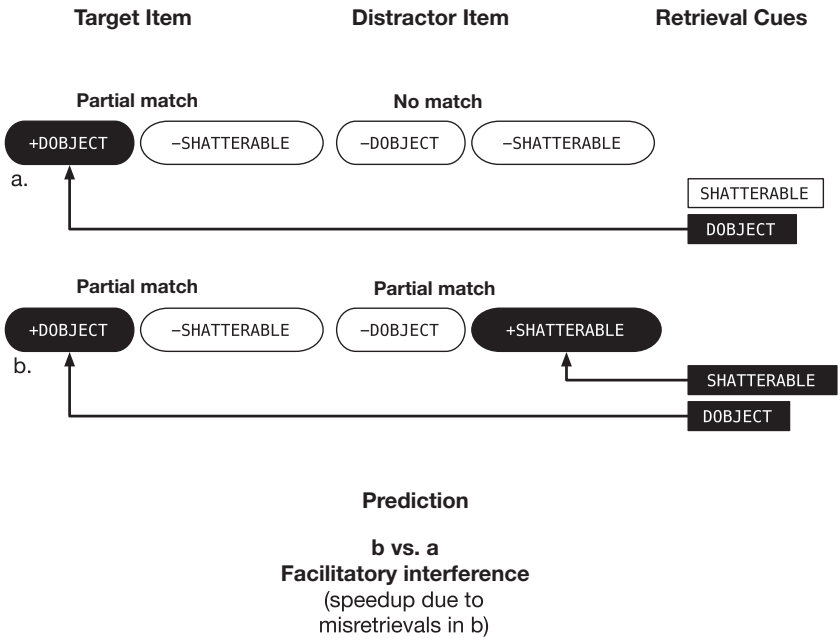

Figure 2.4 Visualization of two conditions in the Cunnings and Sturt (2018) experiment and the predictions of the cue-based retrieval model. The verb *shattered* attempts to retrieve an item in memory that is a direct object and has the property "is shatterable". In both the (a) and (b) conditions shown, the direct object (which is the target noun that should be retrieved) matches the direct object retrieval cue. However, in (b), the distractor noun matches the "is shatterable" cue. As a consequence, in (b), both the target and distractor nouns enter into a race, and whichever item is non-deterministically retrieved is the winner of the race. This race process leads to a faster reading time at the verb *shattered* in (b) vs. (a).

2.4 Subject-Verb Number Agreement

It is well-known that sentences such as (14a) can lead to an illusion of grammaticality. The sentence is ungrammatical because of the lack of number agreement between the subject *key* and the auxiliary *are*. Note that the second noun, *cabinets*, and the auxiliary *are* agree in number, but no syntactic agreement is possible between these two elements.

(14) a. *The key to the cabinets are on the table.

b. *The key to the cabinet are on the table.

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

Many sentence comprehension studies have shown that the illusion has the effect that the auxiliary *are* is read faster in (14a) compared to the equally ungrammatical sentence (14b); in the latter case, the second noun (*cabinet*) is singular and does not agree with the auxiliary in number.

In sentence comprehension, one explanation for the agreement attraction effect is in terms of cue-based retrieval. Wagers et al. (2009) suggested that when the parser encounters the verb, the mismatch between the expected number of the verb and the actual number marking triggers a retrieval process. In the above example, the verb triggers a search for a plural-marked noun that is the subject of the verb. This leads to occasional misretrievals of the only plural marked noun in the sentence, *cabinets*. An obvious problem with this account is that it seems unlikely that the reader interprets the sentence to mean that the cabinets are on the table; of course, such an objection assumes that the reader is engaged in fully interpreting the sentence, which itself may be a questionable assumption (Ferreira et al., 2002; Sanford and Sturt, 2002); we return to the question of underspecification later (Chapter 6). Note that the explanation for subject-verb number agreement conditions is the same as that for Cunnings and Sturt's data for their sentences (13 above). One important difference between the Wagers et al. design and that of Cunnings and Sturt is that in the latter it is very plausible that the reader incorrectly retrieves the distractor as a subject (although Cunnings and Sturt did not check whether readers did in fact misinterpret the sentence). It is not clear whether such a misretrieval occurs in subject-verb number agreement.

Another possible explanation for the agreement attraction effect is in terms of the feature overwriting model of Nairne (1990). In Example (14b), both the nouns are marked singular, whereas in Example (14a) the nouns have different number marking. As discussed in Villata and Franck (2016), the similarity in number of the two nouns in (14b) could be the underlying cause for increased processing difficulty compared to (14a). The identical number marking in (14b) could lead to increased confusability between the two nouns, leading to longer reading times at the moment when a subject noun is to be accessed at the auxiliary verb. The feature overwriting model of Nairne (1990) formalizes this idea. To quote (p. 252): "*An individual feature of a primary memory trace is assumed to be overwritten, with probability F, if that feature is matched in a subsequently occurring event. Interference occurs on a feature-by-feature basis, so that, if feature b matches feature a, the latter will be lost with probability F*." This proposal can be formalized as a hierarchical mixture model (Vasishth et al., 2017a), as we discuss in Section 7.3.

A third explanation for agreement attraction is in terms of the Marking and Morphing (MM) model; this model is intended to explain effects in production rather than comprehension. Under the MM model, attraction arises due to ambiguous encoding of the number marking on a subject phrase (e.g., Eberhard

et al., 2005). In MM, number is considered to be a continuum and not a binary value. The feature "plural" from the distractor noun (i.e., the attractor) spreads activation to the root node of the subject noun phrase, causing it to become more "plural". The extent to which the subject noun phrase becomes plural depends on factors such as the number of distractor nouns with the plural feature, and how near they are to the subject noun phrase's root node in the syntactic tree. Hammerly et al. (2019) provide an implementation of MM that seeks to explain grammaticality judgement data in terms of a drift diffusion process (Ratcliff, 1978). In the Hammerly et al. implementation, the basic explanation for ungrammatical agreement attraction configurations being judged grammatical erroneously is a slower rate of evidence accumulation in favour of the correct and incorrect dependency completion. This model has not yet been extended to explain reading times, and it is not clear whether under this model attraction is limited to retroactive interference designs and not proactive (Avetisyan et al., 2020), but it is an interesting proposal that needs further development.

Related evidence for an encoding account for agreement attraction comes from Paape et al. (2020). This work presents empirical evidence from Eastern Armenian that the number feature percolates to the grammatical subject from both the distractor noun as well as the verb. Paape and colleagues implement a series of competing computational models and show that the model that best explains the data is one that allows feature overwriting on the grammatical subject.

These different theories/explanations for the agreement attraction effect are not necessarily mutually exclusive; any combination of these theories, or possibly all of them, could together explain the data. Such hybrid models have not yet been developed or tested; developing them is an interesting direction for future research.

As shown in Figure 2.5, there is remarkable variability in agreement attraction data, but the posterior distribution of the effect has mean -22 ms, with 95% credible interval $[-37, -9]$ ms, which is consistent with an overall facilitation effect. These estimates are remarkably consistent with the facilitatory effects observed in the two experiments by Cunnings and Sturt (2018) (-22 ms, $[-4, -42]$ ms, and -19 ms, $[1, -40]$ ms). As discussed earlier, Cunnings and Sturt's experiments involved a plausibility manipulation, not the number feature; this could mean that such facilitatory effects are a hallmark of configurations in which the features on the item targeted for retrieval don't fully match all the retrieval cues.

Almost all the data displayed in Figure 2.5 comes from languages like English and Spanish (an exception is Tucker et al., 2015, who investigated Arabic). English and Spanish have relatively impoverished case marking systems. What happens if the grammatical subject and object are unambiguously case-

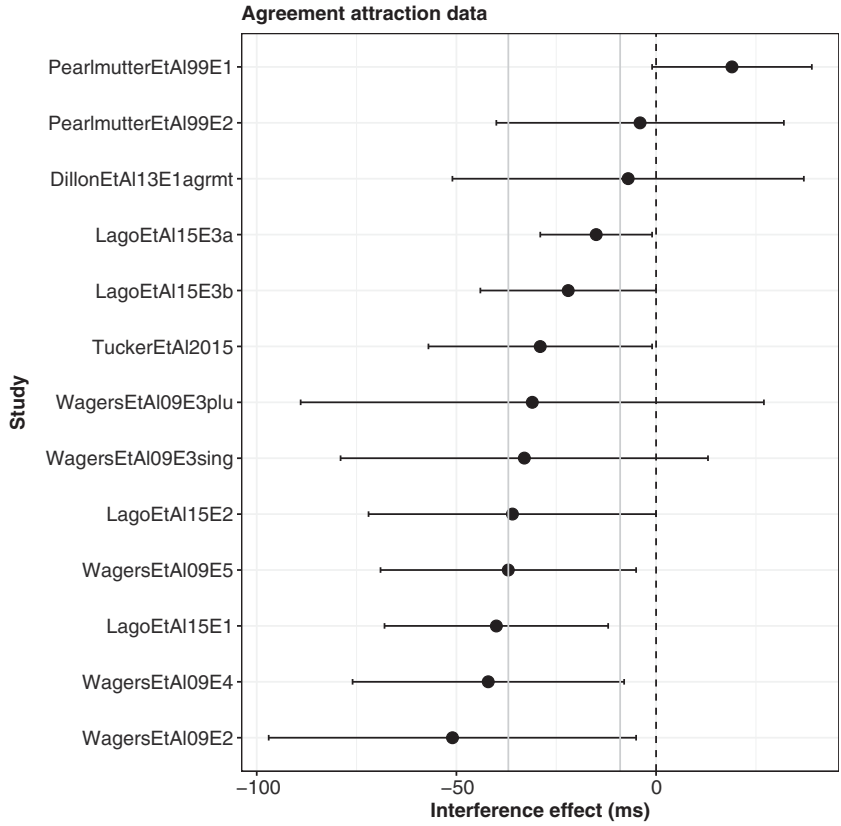

Figure 2.5 Subject-verb number agreement effects in ungrammatical sentences (reading studies). Shown are the means (sorted by increasing magnitude of the effect) and 95% confidence intervals that were either computed from publicly available data or derived from published estimates.

marked? If case marking allows the parsing system to sufficiently distinguish between the nouns, the agreement attraction effect should be weakened when the nouns have distinctive case marking. Avetisyan et al. (2020) tested this hypothesis using Armenian, a language with subject-verb agreement and rich case marking. In a series of experiments (forced choice and self-paced reading), they found that although distinctive case marking on subject and object nouns led to facilitation in processing, there was no indication that distinctive case marking attenuates the agreement attraction effect. One explanation offered by Avetisyan et al. (2020) for the absence of an interaction between case marking and agreement attraction is cast in terms of predictive parsing processes. As

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

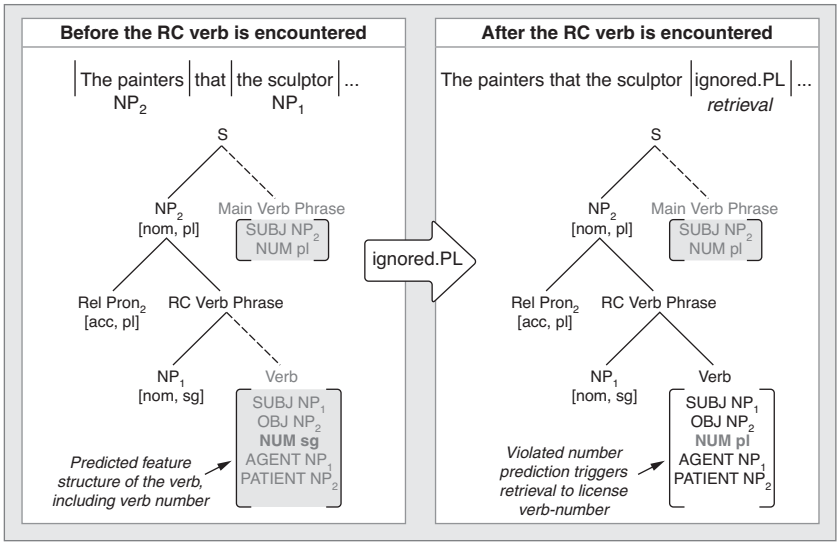

Figure 2.6 The role of case marking in agreement attraction configurations. The figure is reused here under a CC-BY4.0 license and is available from https://doi.org/10.6084/m9.figshare.11440854.v1.

shown schematically in Figure 2.6, once the nouns have been read, the parser predicts a verb phrase with the subject and object subcategorization features already linked to the previously processed nouns. For example, if the reader encounters a sentence like *The painters that the sculptor…*, a singular-marked verb is predicted, but the subcategorization frame of the verb is already filled with the indices corresponding to the subject and object nouns. Now, if a plural-marked verb is encountered, only the number marking of the predicted chunk needs to be modified to integrate the verb with the predicted verb phrase chunk. After that integration, agreement attraction may happen in the manner that Wagers et al. (2009) suggest. If case marking only plays a role during prediction, as suggested above, this may explain why Avetisyan et al. find no indication that distinctive case marking attenuates the agreement attraction effect.

The number attraction examples discussed in Figure 2.6 involve ungrammatical sentences. Grammatical versions of the number agreement configuration have also been investigated. Examples are shown in the following.

(15) a. The keys to the cabinet are on the table.
b. The keys to the cabinets are on the table.

Here, the general claim in the reading literature (Lago et al., 2015) is that no difference is seen between the two conditions. If we examine the estimates

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

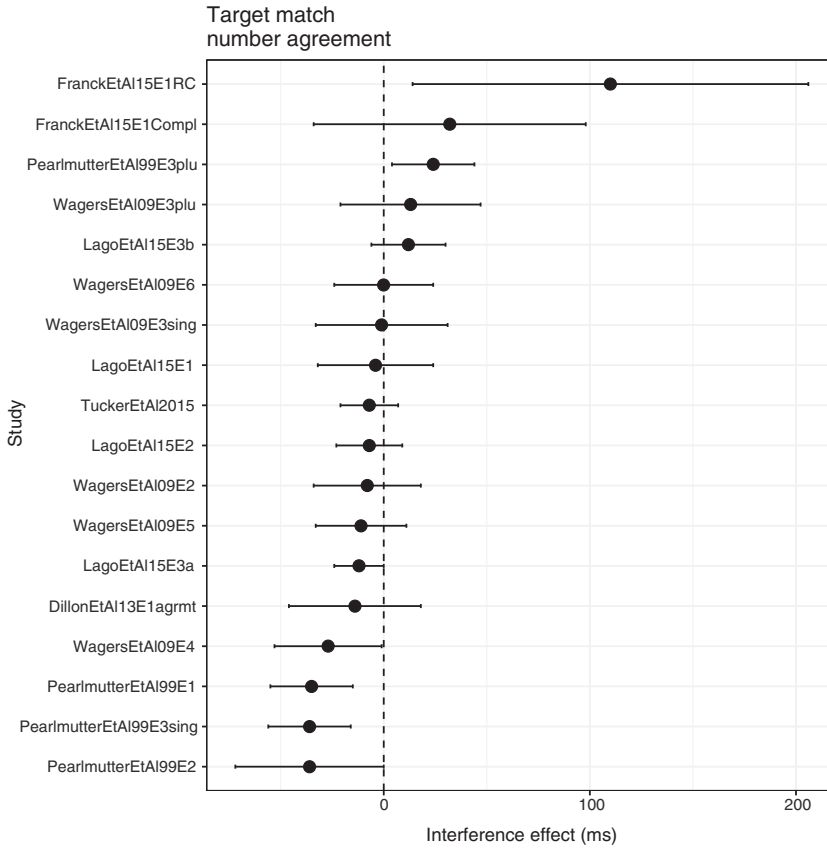

Figure 2.7 Target match number agreement effects in reading studies.

from these studies, we again see a wide range of variability, with all possible outcomes being observed; see Figure 2.7. The mean of the posterior distribution of this effect (the reading time at the auxiliary in (15b) minus the reading time at the auxiliary in (15a) across all these studies (some studies used the post-critical region) is -7 ms, with 95% credible interval $[-17, 4]$ ms.

Based on the studies from their lab, Wagers and colleagues conclude that there is no difference in processing in the two grammatical agreement attraction conditions shown in (15a, 15b). Wagers et al. explain this null effect as follows. The subject noun predicts a verb with a particular number marking, and this prediction is validated when the verb is encountered. In such a situation, no retrieval process is triggered. This proposal has some difficulties. A great deal of work on English (Bartek et al., 2011; Gibson, 2000; Grodner and Gibson, 2005) has consistently shown that even in grammatical constructions, a retrieval

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

process is triggered. It seems implausible that retrieval is not triggered only in this one particular case, where the number feature is involved.

How strong is the evidence for the null results reported in the Dillon et al., Wagers et al., and Lago et al. studies? When *p*-values are greater than 0.05, this is not necessarily evidence that the null hypothesis is true. As discussed in Section 1.5.4, when power is low, it is hardly surprising that repeated experiments show null results. This point has somehow been lost in the course of translating statistical theory to psychological and linguistic applications. Instead of concluding that they have no evidence for an effect, researchers will incorrectly conclude that "absence of evidence is evidence of absence."

What would have happened if statistical power were higher than in the studies mentioned above? The Dillon et al., Wagers et al., and Lago et al. studies generally have small sample sizes, leading to power far below 80%. Nicenboim et al. (2018) increased power by increasing sample size to 185 subjects and by increasing the strength of the interference manipulation. Their design involved grammatical German sentences with number interference. Here, a subject noun and a verb always have two nouns intervening between them. In the high-interference condition, all three nouns match the number feature that is on the verb; in the low-interference condition, only the subject noun has the number feature that is on the verb. Thus, this design seeks to increase the magnitude of the number interference effect by increasing the fan, that is, by increasing the number of nouns that match the retrieval cues.

(16) a. High Interference

Der Wohltäter, der den Assistenten des Direktors **begrüsst hatte**, sass später im Spendenausschuss.
The.sg.nom philanthropist, who.sg.nom the.sg.acc assistant (of) the.sg.gen director **greeted had.sg**, sat.sg later in the donations committee.

'The philanthropist, who had greeted the assistant of the director, sat later in the donations committee.'

b. Low Interference

Der Wohltäter, der die Assistenten der Direktoren **begrüsst hatte**, sass später im Spendenausschuss.
The.sg.nom philanthropist, who.sg.nom the.pl.acc assistant(s) (of) the.pl.gen director(s) **greeted had.sg**, sat.sg later in the donations committee.

'The philanthropist, who had greeted the assistants of the directors, sat later in the donations committee.'

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

This larger-sample study suggests that the magnitude of the cue-based retrieval effect in grammatical sentences involving number agreement may be smaller compared to the effect observed in ungrammatical agreement attraction configurations. Nicenboim et al. demonstrate that if the number of distractor nouns is increased from one to two, a small interference effect can be observed at the verb *begrüsst hatte*, 'greeted had', in sentences like (16a) compared to (16b). The authors found that with two distractors present, the interference effect is approximately 9 ms with a 95% credible interval of 0–18 ms. What could be the reason for the smaller interference effect in this case? Nicenboim et al. argue that feature percolation (the mechanism assumed in the MM model) and cue-based retrieval may be acting in opposite directions. It follows that if one increases the magnitude of the interference effect, the effect should be detectable. This proposal has yet to be tested with new experimental designs and is an interesting avenue for future research.

2.5 Reflexives and Reciprocals

Sturt (2003) carried out an eyetracking study that investigated the processing of direct object reflexives. He suggested that when the parser encounters a reflexive, in the first moments of processing, the antecedent is chosen using principle A of the binding theory (Chomsky, 1981). This implies that if any other noun phrases are present that are not syntactically licensed as antecedents of the reflexive, these would never be considered as possible antecedents even if the gender marking on the reflexive matches these noun phrases. Two examples are shown below to illustrate the two basic configurations that have been studied in the literature. These examples are adapted from Sturt's paper.

(17) a. Proactive
Jonathan/Jennifer remembered that the surgeon had pricked himself with a used syringe needle.

b. Retroactive
The surgeon who Jonathan/Jennifer met had pricked himself with a used syringe needle.

Example (17a) shows a proactive interference configuration: the reflexive *himself* requires the subject of the local clause, that is, *surgeon*, as the legal antecedent. However, the proper noun *Jonathan* matches in gender with the reflexive. Under the Sturt account, in the first moments of processing, compared to the baseline where the distractor noun (e.g., *Jennifer*) doesn't match the gender of the reflexive *himself*, the masculine-marked distractor noun *Jonathan* would never be considered as an antecedent. Example (17b) shows a retroactive interference configuration: the distractor noun *Jonathan* appears between the subject, which is the antecedent of the reflexive, and the reflexive *himself*.

In both configurations, one can investigate the effect of the distractor noun by comparing sentences that either have a masculine distractor noun such as *Jonathan*, or a feminine distractor noun such as *Jennifer*. Sturt found no evidence that the reflexive was mistakenly associated with the distractor noun at the earliest moments of processing, that is, in first-pass reading times. As Sturt puts it (p. 542), "Principle A of the binding theory operates at the very earliest stages of processing; … the gender of the ungrammatical antecedent [the distractor noun] had no effect on early processing, although it affected processing during later stages." In other words, at the earliest moments of processing, based on these null results, reflexives are assumed to be immune to the effects of interference.

Recall that earlier we had seen in subject-verb dependencies that interference effects are robustly seen. In the grammatical subject-verb dependencies investigated by Van Dyke and colleagues, we robustly see inhibitory effects, and in ungrammatical subject-verb dependencies with number agreement between the distractor and verb, we see a relatively clear indication of facilitation effects. Since the majority of these studies involve self-paced reading, we cannot say whether these inhibitory and facilitatory effects reflect the earliest moments of processing. An exception is the eyetracking study by Van Dyke (2007); but here, too, first-pass reading time seems to show no interference effects (see Figure 2.2). However, in a recent larger-sample eyetracking study involving English, Mertzen et al. (2020a) did find the predicted inhibitory interference effects in first-pass reading times.

Is the processing of reflexives different from those of subject-verb constructions? The answer would be yes if interference effect was seen in subject-verb agreement constructions but not in reflexive constructions. In particular, at the earliest moments of processing (e.g., in first-pass reading times), we would expect to see interference effects in subject-verb constructions but not in reflexives. Dillon et al. (2013) were the first to directly compare interference effects in these two dependency types (their experiment 1). They compared subject-verb number agreement constructions with reflexives. See (18, 19).

Number agreement conditions:

(18) a. Grammatical
The new executive who oversaw **the middle manager** apparently **was** dishonest about the company's profits

b. Grammatical
The new executive who oversaw the middle managers apparently was dishonest about the company's profits

c. Ungrammatical
*The new executive who oversaw the middle manager apparently **were** dishonest about the company's profits

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

d. Ungrammatical
*The new executive who oversaw **the middle managers** apparently **were** dishonest about the company's profits

Reflexive conditions:

(19) a. Grammatical
The new executive who oversaw **the middle manager** apparently doubted **himself** on most major decisions

b. Grammatical
The new executive who oversaw the middle managers apparently doubted **himself** on most major decisions

c. Ungrammatical
*The new executive who oversaw the middle manager apparently doubted **themselves** on most major decisions

d. Ungrammatical
*The new executive who oversaw **the middle managers** apparently doubted **themselves** on most major decisions

Dillon generously gave us the data from his study. This allowed us to determine whether, in early vs. late measures, any difference is seen between agreement and reflexives. We first defined nested contrasts (in grammatical and ungrammatical sentences separately) as shown in Table 2.1; for more details on contrast coding, see Schad et al. (2020b). Note that Dillon and colleagues used a different contrast coding than we did (main effects and interactions of grammaticality and intrusion); the details of these differences are discussed in Jäger et al. (2020). Our contrast coding was designed to directly test the predictions of the Lewis and Vasishth (2005) model.

We analyzed all dependent measures that have been invoked as indexing early processes in the dependencies considered in this chapter: first-pass reading time and regression probability (Dillon et al., 2013), and regression path duration (Cunnings and Sturt, 2018). As shown in Figure 2.8, the only clear effect is in total reading times in ungrammatical agreement dependencies. None of the dependent measures that is claimed to index early processes shows any effects in either agreement or reflexive dependencies. Thus, from these data at least, there is no reason to believe that interference effects *ever* occur in early measures in *any* dependency, as claimed by Sturt (2003). It is therefore not clear why reflexive processing is seen as special and different from any other dependency. One could conclude that all dependencies uniformly show an absence of interference effects in early measures.

In their paper, Dillon and colleagues argue that reflexives and agreement attraction constructions exhibit different interference profiles in ungrammatical constructions. In order to argue for a difference between agreement and

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

Table 2.1. *Nested contrast coding to investigate the effect of intrusion in grammatical and ungrammatical agreement and reflexive constructions. The contrast* dep *is the main effect of dependency type (agreement or reflexive). The abbreviation* intr.au *means intrusion (interference effect) in agreement dependencies, ungrammatical;* intr.ag *stands for intrusion (interference effect) in agreement dependencies, grammatical;* intr.ru *refers to intrusion (interference effect) in reflexive dependencies, ungrammatical;* intr.rg *stands for intrusion (interference effect) in reflexive dependencies, grammatical.*

	Agreement				Reflexives			
	Gram		Ungram		Gram		Ungram	
	No intr	Intr	No intr	Intr	No intr	Intr	No intr	Intr
dep	–0.5	–0.5	–0.5	–0.5	0.5	0.5	0.5	0.5
intr.au	0	0	–0.5	0.5	0	0	0	0
intr.ag	–0.5	0.5	0	0	0	0	0	0
intr.ru	0	0	0	0	0	0	–0.5	0.5
intr.rg	0	0	0	0	–0.5	0.5	0	0

reflexives with respect to the interference manipulation, an interaction must be demonstrated between dependency type and the interference manipulation. However, such an interaction was not present in the original data (Jäger et al., 2020).

Thus, although the experiment design had the potential to demonstrate that dependency type determines whether interference occurs, the data don't seem to provide a basis for a conclusion.

A major issue in the Dillon et al. study was that the sample size was quite small. A prospective power analysis of the Dillon et al. data – using their sample size and predicted effects from the cue-based retrieval model of Lewis and Vasishth – shows that prospective power is likely to have ranged from 20% to 40% for subject-verb agreement configurations and 5–25% for reflexives; see Jäger et al. (2020) for details on the power calculations.

We attempted to replicate the key results from total reading times with a larger sample size (181 participants). This work is reported in full in Jäger et al. (2020). Figure 2.9 shows the total reading times at the critical region (the auxiliary or reflexive). Figure 2.9 shows that both agreement and reflexives in ungrammatical conditions show very similar facilitatory interference patterns in total reading times. These similar estimates for the two dependencies are not consistent with the claims by Dillon and colleagues. However, an exploratory analysis of first-pass regressions did show some weak evidence consistent with the interaction pattern predicted by Sturt and Dillon and colleagues. If this

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

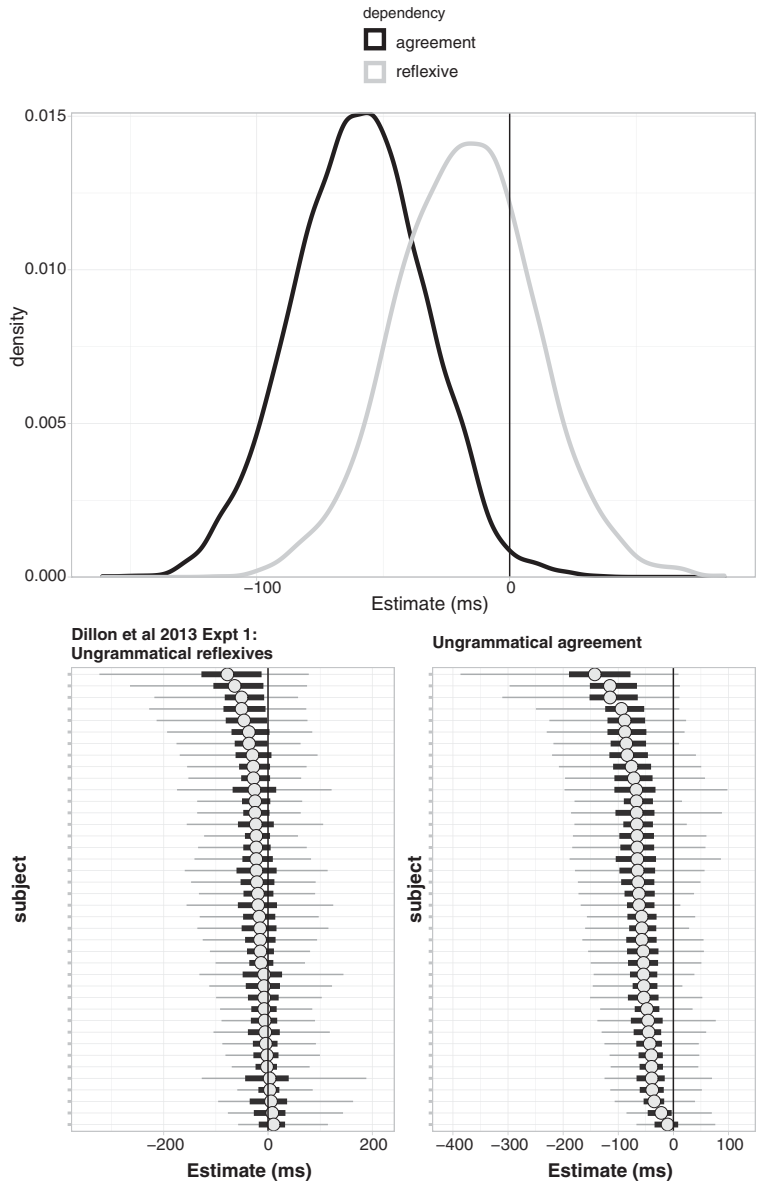

Figure 2.8 Summary for total reading time of the Dillon et al. (2013) comparisons for ungrammatical sentences involving agreement and reflexives. The sample size was 40 participants. The upper plot shows the posterior distributions of the facilitatory interference effect in agreement and reflexives, and the lower plots show the individual-level estimates of the effect, with 80% and 95% credible intervals.

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

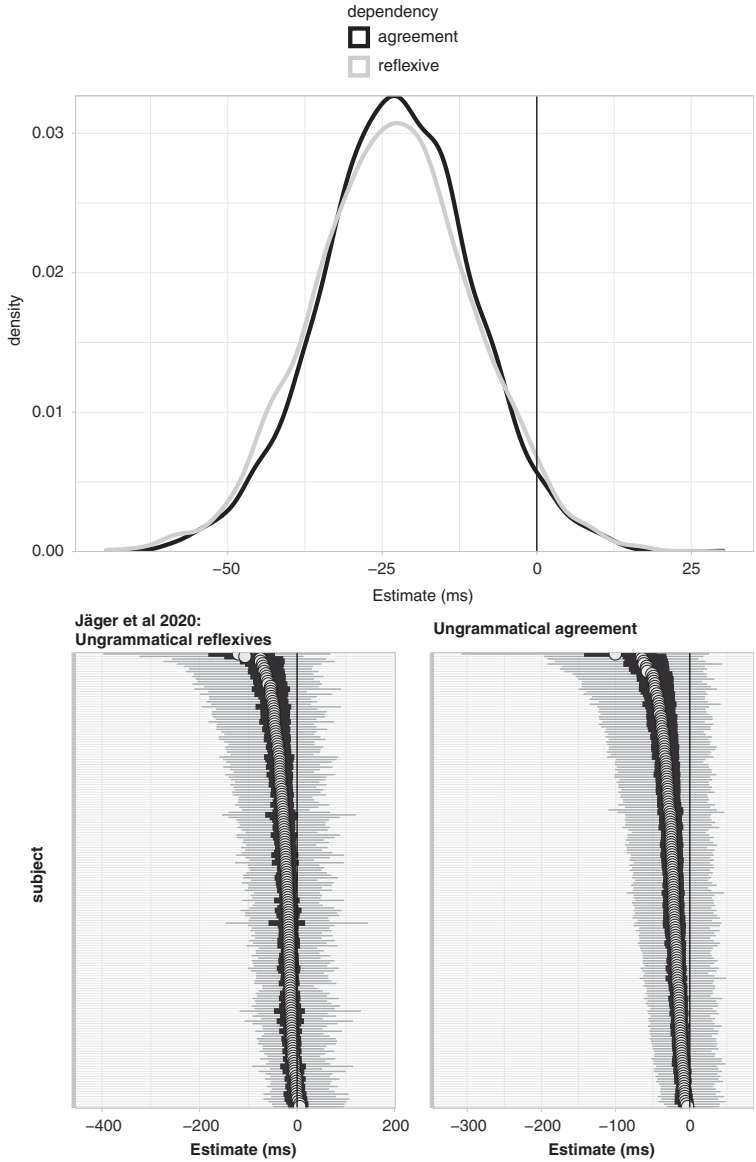

Figure 2.9 Summary for total reading time of the Jäger et al. (2020) comparisons for ungrammatical sentences involving agreement and reflexives. The sample size was 181 participants. The upper plot shows the posterior distributions of the facilitatory interference effect in agreement and reflexives, and the lower plots show the individual-level estimates of the effect, with 80% and 95% credible intervals.

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

pattern can be replicated in future work, it would suggest that Sturt's original proposal may have been correct, that only in the early moments of processing (expressed in first-pass regressions) is there immunity from interference; in later stages (expressed by total reading times), agreement and reflexive dependency types show similar interference profiles.

As an aside, it is worth noting here that if the Jäger et al. (2020) study's results for total reading times had been interpreted using the strict cut-off of 0.05 for the *p*-value practiced in psycholinguistics, we would be forced to conclude that there is no effect of agreement or reflexives! This kind of simplistic conclusion is a good example of why the *p*-value-based decision process that is standardly used in psycholinguistics is deeply flawed.

2.5.1 *Individual-Level Effects in the Dillon et al. Design*

Figures 2.8 and 2.9 show an interesting consistency across the original Dillon et al. study and the Jäger et al. replication attempt: essentially all the subjects show facilitatory interference effects in both agreement and reflexive constructions, in both the original experiment and the replication attempt. In both studies, the magnitude of the effect varies in the two dependencies from subject to subject, but the sign is consistently negative. This is a potentially interesting pattern that could have a theoretical explanation. For example, some subjects might show very large facilitatory interference effects because they are engaged in good-enough processing, or are not using syntactic constraints to complete dependencies to the same extent as other subjects, who show smaller facilitatory interference effects. This modulation of the effect size for individual subjects can be modelled in the Lewis and Vasishth (2005) architecture, as we show in Section 3.2.1, and in Yadav et al. (2020).

So, given the above data (the original Dillon et al. data and our replication data), what should we conclude about the processing of reflexives and subject-verb agreement? In psycholinguistics, researchers feel obliged to take one or the other position. This type of deterministic thinking is highly misleading. A more realistic approach is to simply lay out what we learnt from the data given different prior beliefs about the problem. We discuss this point next.

2.5.2 *A Sensitivity Analysis on the Ungrammatical Agreement and Reflexives Conditions Using Informative Priors*

In this chapter, we summarized the main evidence available from reading studies relating to interference effects in different dependency types. On the surface, one might think that empirical data are "objective" in the sense that they speak for themselves. However, in practice researchers always interpret data in the light of their prior beliefs, and sometimes these beliefs can be

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

very strong. When these prior beliefs are strong, it makes intuitive sense that a single counterexample from one experimental result should not change our beliefs much. In the case of reflexives, in the course of informal discussions, researchers have expressed scepticism about the estimates of the facilitatory interference effect in English reflexives reported in Jäger et al. (2020). The argument here is that there is a lot of prior data that doesn't match the Jäger et al. findings. Essentially, the objection is that in the analysis of the Jäger et al. replication data, we do not use all available information from prior work. This kind of use of prior knowledge need not be invoked informally; one can simply take one's prior beliefs into account in the data analysis. This is not normally done in psycholinguistics, but with the increasing availability of Bayesian tools for data analysis, it is easy to formally incorporate prior beliefs into account.

In this section, we briefly demonstrate how prior beliefs can be taken into account in the specific case of the data from Jäger et al. Below, we illustrate how the expert can formally interpret available data in the light of either prior data or their own prior subjective beliefs. The reason we bring this point up here is that we feel that incorporating prior beliefs in the analysis is a very important tool for understanding "what the data tell us". The data never "speak for themselves"; they always speak through the filter of our beliefs. We show below how this subjectivity can be formally incorporated into the interpretation of data.

All the statistical models fit in Jäger et al. (2020) used mildly uninformative, regularizing priors, which effectively assume an agnostic starting point (Schad et al., 2020a). Bayesian methods allow us to quantitatively take into consideration prior knowledge or beliefs about the plausible values of a parameter, by using an informative prior.

Priors may arise from different sources, an obvious one being a body of empirical data on the specific issue in question. When empirical data are scarce, other sources can be expert opinion (Oakley and O'Hagan, 2010; O'Hagan et al., 2006), or via quantitative predictions of existing theories.

Speaking informally, the posterior mean can be seen as a weighted sum of the prior mean and the sample mean, weighted by the relative precision (inverse of the variance) of the prior and the data. If the prior has relatively higher precision, it will dominate in determining the posterior mean, and if the data has higher precision (this is a function of standard deviation and the sample size), then the data will dominate in determining the posterior mean. A consequence of this fact is that when we have a very strong prior belief, expressed through a distribution with a relatively small standard deviation, even large amounts of data may not shift the posterior mean away from the prior mean.

Hence, when investigating a controversial research question, it may be desirable to quantitatively take into account opposing theoretical views by using a representative spectrum of different priors. In this context, medical statisticians

like Spiegelhalter et al. (2004) have proposed the use of a "community of priors": opposing perspectives of researchers are incorporated in the data analysis by using different priors. In this way, one can use *agnostic priors* (mildly uninformative priors), *enthusiastic priors* that support a particular position, and *adversarial or sceptical priors* that represent alternative positions. The different posterior distributions from the data can then be examined in the light of these priors, and the researcher can draw their own conclusion.

In the reflexives replication data, we can examine next how agnostic, adversarial, and enthusiastic priors affect the posterior distributions of the effects of interest. The case of ungrammatical agreement constructions is relatively uncontroversial, and we will see below that a range of different priors have little effect on the estimates from the Jäger et al. data. More interesting is the effect of different prior specifications on the interference effects in ungrammatical reflexive conditions. We demonstrate here that the effect estimates here are quite sensitive to prior beliefs.

We carried out a sensitivity analysis on the estimates for the ungrammatical conditions, defining priors that represent three sources of beliefs. The different priors are summarized in Table 2.2.

(i) **Mildly uninformative priors (Agnostic prior)** As a baseline, we used a mildly uninformative prior for both the agreement and reflexive conditions. This prior represents an agnostic starting point where no information is incorporated from any prior knowledge.

(ii) **Meta-analysis priors (Adversarial prior for reflexives)** We derived posterior distributions of the interference effect in ungrammatical agreement and reflexive conditions using data from existing reading time studies (Jäger et al., 2017). These studies represent a synthesis of the evidence available from self-paced reading and eyetracking studies on agreement and reflexives. Because the dependent measure of interest in our studies is total fixation time, the estimates from the eyetracking studies are based on total fixation times. We refer to this prior as an adversarial prior because the estimate for ungrammatical reflexives is $N(9,10.75)$, which is a relatively tight prior for the reflexives interference effect in our replication study. A great deal of data would be needed to shift the posterior mean such that a facilitatory interference effect is seen.

(iii) **LV05 priors (Enthusiastic prior)** As a prior representing the equal cue-weighting retrieval proposal, we used a normal approximation of the range of predicted effects from the Engelmann et al. (2020) model.

The results of this sensitivity analysis are shown in Table 2.2. Agreement conditions show similar facilitatory interference effects regardless of the prior chosen. This confirms that the agreement interference effect is robust to the choice of prior. For reflexives, the situation is different. The reflexive conditions

Table 2.2. *Summary of the sensitivity analysis investigating the effect of incorporating prior knowledge from mildly uninformative priors; a meta-analysis of existing reading data on ungrammatical agreement and reflexives; and the model predictions in Engelmann et al. (2020). The dependent measure in the analysis is total fixation time and the posterior estimates are back-transformed to the ms scale from log ms. The priors are shown in the ms scales.*

Sensitivity analysis			
Condition	Source for Prior	Prior (ms)	Posterior (ms)
Agreement (Ungram)	Mildly uninformative	N(0,7600)	–22 [–46,1]
	Meta-analysis	N(–32,8.5)	–25 [–36,–14]
	LV05 Model	N(–26,13)	–22 [–34,–7]
Reflexives (Ungram)	Mildly uninformative	N(0,7600)	–24 [–50,2]
	Meta-analysis	N(9,10.75)	–3 [–17,12]
	LV05 Model	N(–26,13)	–22 [–38,–6]

show facilitatory interference effects only when we use mildly uninformative priors and the LV05 predictions as priors. With the relatively tight meta-analysis prior $N(9,10.75)$, we see no indication of facilitatory interference effects in the replication data.

Thus, for reflexives, the conclusion that one can draw from these data is not as clear as for agreement; the conclusion depends quite a bit on the researcher's prior beliefs. Of course, future studies could now use the Jäger et al. reflexives estimates as priors. By incrementally incorporating prior knowledge in newer and newer studies, eventually it could become clear what the facts are about reflexives.

2.6 Concluding Remarks

The reading studies that have investigated these different types of dependency constructions show very limited evidence for inhibitory and facilitatory interference. Although subject-verb non-agreement dependencies seem to often show patterns consistent with inhibitory interference, grammatical subject-verb dependencies often do not. This is a puzzle for similarity-based interference researchers. Furthermore, it has been argued that reflexives (and perhaps also reciprocals) are immune to inhibitory or facilitatory interference because binding theory ensures that the antecedent is unerringly found in memory. If this claim turns out to be true, it would be an interesting and important exception to the general principles of interference that are claimed to apply in sentence processing. One important implication would be that linguistic

https://doi.org/10.1017/9781316459560.007 Published online by Cambridge University Press

constraints, often subtle constraints, can play a greater role in online processing than general working memory constraints. This would show, inter alia, that sentence processing is at least partly subject to purely linguistic constraints.

One thing that stands out from reviewing the published evidence (Jäger et al., 2017) is the generally low statistical power of the published studies. If the meta-analytical estimates of inhibitory and facilitatory interference effects are accurate estimates of the true underlying effects, then none of the published reading studies so far can be considered properly powered for detecting the effects. The reason: the effects are too small to be capable of being detected accurately by the published studies.

The single biggest reason that low-power studies have proliferated in psycholinguistics is that fundamental misunderstandings exist in psycholinguists' minds about what a *p*-value can and cannot not tell us, given no other information. Null results are widely assumed to indicate that the true effect is 0, and statistically significant, exaggerated estimates, which never replicate, get published as big-news results. We discuss these points at length in Vasishth et al. (2018), Nicenboim et al. (2018, 2020), and Jäger et al. (2020).

This is a disappointing empirical starting point for evaluating the computational models considered in this book. As discussed in the previous chapter, one of the Roberts and Pashler (2000) criteria for a persuasive model fit – higher precision data – has not yet been met in the literature. But the situation is what it is. Given the data that are available today, it is possible to draw some initial conclusions about the models' performance. That is what the rest of this book tries to achieve. Our hope is that some day there will be higher-precision benchmark data for evaluating sentence processing models of the sort discussed here.