GLOSSA psycholinguistics

Schmitz, T., Winkowski, J., Hoeks, M., Nouwen, R., & Dotlačil, J. (2024). Semantic accessibility and interference in pronoun resolution. *Glossa Psycholinguistics*, 3(1): 1, pp. 1–58.
DOI: https://doi.org/10.5070/G60111424

Semantic accessibility and interference in pronoun resolution

Tijn Schmitz, Institute for Language Sciences, Utrecht University, NL, t.p.a.schmitz@uu.nl
Jan Winkowski, Institute for Language Sciences, Utrecht University, NL, j.l.winkowski@uu.nl
Morwenna Hoeks, Department of Linguistics, University of California, Santa Cruz, US, morwennahoeks@gmail.com
Rick Nouwen, Institute for Language Sciences, Utrecht University, NL, r.w.f.nouwen@uu.nl
Jakub Dotlačil, Institute for Language Sciences, Utrecht University, NL, j.dotlacil@uu.nl

The general view in the syntactic literature is that binding constraints can make antecedents syntactically inaccessible. However, several studies showed that antecedents which are ruled out by syntactic binding constraints still influence online processing of anaphora in some stages, suggesting that a cue-based retrieval mechanism plays a role during anaphora resolution. As in the syntactic literature, in semantic accounts like Discourse Representation Theory (DRT), formal constraints are formulated in terms of accessibility of the antecedent. We explore the discourse inaccessibility postulated in DRT by looking at its role in pronoun resolution of inter-sentential anaphoric relations in four off-line and two eye-tracking experiments. The results of the eye-tracking experiments suggest that accessibility has an effect on pronoun resolution from early on. The study quantifies evidence of inaccessible antecedents affecting pronoun resolution and shows that almost all evidence points to the conclusion that discourse-inaccessible antecedents are ruled out for pronoun resolution in processing. The only potential counter-example to this claim that we detected remains only as weak evidence, even after combining data from both eye-tracking studies. The findings in the study show that accessibility plays a significant role in the processing of pronoun resolution, in a way which is potentially challenging for the cue-based retrieval mechanism. The paper argues that discourse accessibility can help expand the theories of retrieval beyond the syntactic and sentence-level domain and provides a window into the study of interference in discourse.

1. Introduction

Sentence comprehension and production undoubtedly rely on memory, and nowhere is this reliance more visible than in the case of the comprehension of so-called *dependents*: elements whose interpretation and/or form depends on another linguistic item. An example of a dependent is the verb *talks* in (1), whose form depends on the morphological specification of the subject *John*. The verb also depends on the subject for its interpretation, in particular, the subject specifies that the referent John fills in one of the roles of the verb, the agent. Thus, to arrive at the correct form and interpretation of the verb, one has to recall what the subject is when producing or comprehending the verb.

(1) John always talks about Mary.

One prominent line of research argues that the recall of the dependent happens via a cue-based retrieval mechanism: it accesses the items in memory that match the target in retrieval cues (Lewis et al., 2006; McElree, 2000; McElree et al., 2003; Nicenboim & Vasishth, 2018; Van Dyke, 2007; Van Dyke & Lewis, 2003; Vasishth et al., 2008; Wagers et al., 2009). In (1), there are several features that could guide the retrieval. For example, the item has to be *singular*, *3rd person* and *subject*. The item also has to be *human*, since only humans talk. All of these features are present on the proper noun *John* and therefore, they can be used to access it.

While cue-based retrieval became well accepted and maybe even the default model in the case of agreement resolution and thematic integration, as in (1), and has been considered for other intra-clausal dependencies (Cunnings & Sturt, 2018; Dillon et al., 2013; Jäger et al., 2017, 2020; Kush & Phillips, 2014; Parker, 2022), it is far from clear that it can or should be used to model all types of linguistic dependencies. In particular, very little is known about dependencies that operate across a discourse. In this paper, we investigate the role of memory retrieval for pronoun resolution in short (up to three sentence-long) discourses. Using the eye-tracking-while-reading paradigm, we show that the resolution of these pronouns crucially differs from previously studied dependencies in that the presence of partially matching, but inaccessible, elements does not affect these dependencies – not even at early stages of processing. In particular, we show that non-referentially quantified noun phrases (e.g. *no girl* vs. *no boy*) do not interfere with the search for potential antecedents of a pronoun in a subsequent sentence. We also detect one possible counterexample to this claim and quantify the amount of evidence that this counterexample represents.

The remainder of this paper is structured as follows. In the next section (Section 2) we summarize the main facts about cue-based retrieval and the research that studied pronominal dependencies from the perspective of cue-based retrieval. In that section, we also present

Discourse Representation Theory (Kamp, 1981; Kamp & Reyle, 1993), whose findings are relevant for testing cue-based retrieval during pronoun resolution in discourses. We present two eye-tracking experiments testing pronoun resolution in Sections 3 and 4, along with resolution and acceptability judgement tasks, and analyze their combined results in Section 5. We summarize our main findings in Section 6.

2. Cue-based retrieval and anaphoric resolution

Cue-based retrieval in dependencies has been implemented in two models (cf. Nicenboim & Vasishth, 2018): the direct-access model, see McElree (2000), and the activation-based model, see Lewis & Vasishth (2005). Disregarding the implementation details of the models, we will focus on three assumptions that have driven the research in the study of dependency and memory:

(i) *Cue-based retrieval is error-prone*: When encountering the dependent, what might be recalled is a distractor, which is an item in memory that does not fully match the cues triggering the retrieval.

(ii) *Recall is cue-dependent*: The likelihood that the distractor is erroneously recalled increases with the number of features in which the distractor matches the right cues.

(iii) *Retrieval time is sensitive to features (activation-based model only)*: The speed of the recall decreases with the number of features. The speed of the recall is also slower the more the feature is shared across items.

We will show how these assumptions work on the paradigm in (2). The examples in (2-a) and (2-b) are based on a study in Van Dyke (2007), but simplified. The examples in (2-c) and (2-d) come from Wagers et al. (2009) and are based on Pearlmutter et al. (1999).

(2) a. The **resident** who was living near the dangerous neighbor *was complaining* about the investigation.
b. The **resident** who was living near the dangerous warehouse *was complaining* about the investigation.
c. *The **key** to the cell unsurprisingly *were rusty* from many years of disuse.
d. *The **key** to the cells unsurprisingly *were rusty* from many years of disuse.

In (2), the verb-subject dependency requires readers to recall the subject (boldfaced in the example) when reading the verb phrase (italicized in the example). We call the subject phrase a *target* in this example. The other noun phrases present in the clause are *distractors*. In the discussion we will focus on the distractors *dangerous neighbor/warehouse* for (2-a) and (2-b) and *the cell/cells* for (2-c) and (2-d).

In (2-a) and (2-b), the features that are relevant for our discussion are *human* (since humans, unlike, e.g., things or buildings, can normally complain and thus, this feature can guide retrieval) and *subject*. The target (*the resident*) fully matches in these features. The distractor, on the other hand, matches the feature *human* in (2-a) but not in (2-b). Thus, it is expected that the distractor is more likely erroneously recalled in (2-a) than (2-b) (assumptions (i) and (ii)). If this erroneous recall is corrected at least in some instances, slowdown due to repair is predicted in (2-a) compared to (2-b). Slowdown in the case of correct recall is predicted for (2-a) compared to (2-b) also because the feature *human* is shared between the target and the distractor, which should lead to slower retrieval of the target (assumption (iii)).

In (2-c) and (2-d), the subject mismatches the *number* feature and, consequently, the sentence is ungrammatical. Crucially, in (2-d), the distractor *the cells* matches the dependent in number and thus it is more likely to be erroneously recalled compared to (2-c) (assumptions (i) and (ii)). Because of the assumption (iii), this will lead to faster processing in (2-d) compared to (2-c).

The predictions of the models have been to a large extent confirmed. It has repeatedly been found that ungrammatical sentences with a plural verb are read faster when the distractor is also plural, i.e., (2-d) leads to faster reading times than (2-c) (Dillon et al., 2013; Jäger et al., 2017, 2020; Lago et al., 2015; Tucker et al., 2015; Villata et al., 2018; Wagers et al., 2009). The slowdown, predicted for (2-b) compared to (2-a), has also been observed (Jäger et al., 2017; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006; Van Dyke, 2007).

One limitation that cue-based retrieval models have is that it is hard to extend them to cases in which relational information plays a role in establishing a dependency (Kush et al., 2015). Consider (3). One dependency in this example is between the reflexive *himself*, which is dependent in its form and interpretation on the subject *John*.

(3) John should tell Bill about himself.

Three constraints are at work in establishing the antecedent of the reflexive: In this case, the antecedent has to carry the features *masculine* and *singular*, and it must be in a position that *c-commands* the reflexive in its local domain (following the locality constraint of Principle A; see Chomsky, 1981; Reinhart & Reuland, 1993). A phrase X c-commands a phrase Y if and only if Y is contained within X's sister. Therefore, in the syntactic structure of (3), as schematized in **Figure 1**, the subject *John* c-commands both the object *Bill* and the reflexive, because both *Bill* and *himself* are contained within the sister of *John*. Note that the object *Bill* also c-commands the reflexive. In this particular case, it is therefore predicted that the reflexive can have either *John* or *Bill* as its antecedent.

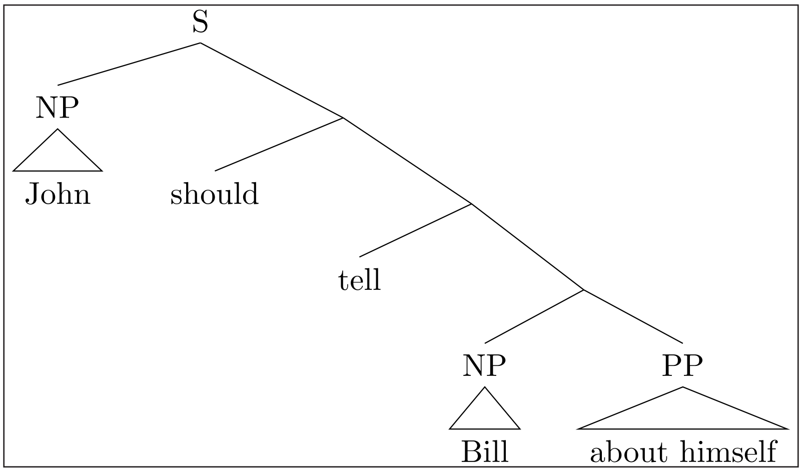


Figure 1: Hierarchical structure for the sentence (3).

The problem that c-command poses for a theory of dependency resolution that is merely based on cue-based retrieval is that c-command is a relational notion and, thus, it cannot be inherent to any element – unlike features like gender or number or being human. To see that, note that while information such as gender or number on an element exists irrespective of what the dependent is, the property of c-commanding a dependent always relies on the position of that dependent. For example, direct objects c-command adjunct PPs, but they do not c-command the subject of the clause. In a right-branching structure, such as the sentence in **Figure 1**, the first word (*John* in this case) c-commands all other terminal nodes in the tree, but this will not be known until after all these terminal nodes are encountered. To fully specify the c-command relation as a feature, we would have to assume that the value of the c-command feature of *John* gets updated every time a new word is read. The fact that the interpretation of reflexives is dependent on such a relational notion poses a challenge for cue-based retrieval, because cue-based retrieval requires the dependent to trigger retrieval based on cues (for a more elaborate discussion, also see Chapter 2 in Kush, 2013). As Kush points out, the problem with this is not that such a feature – and therefore such a cue – cannot be defined. Instead, the problem is that incorporating such a feature would put unrealistic demands on the parser, as it would entail that each time a new word is read, a retrieval has to be launched to update the c-command feature of all preceding items.

It has been argued that structural features like c-command should not be modeled using cue-based retrieval or that such features should have a different status than inherent features; for example, they pre-filter items before cue-based retrieval can take place or they strongly dominate inherent features during cue-based retrieval (see, e.g., Cunnings & Sturt, 2014; Dillon et al., 2013; Van Dyke & McElree, 2011). This position was further strengthened by empirical findings suggesting that non c-commanding distractors do not affect the resolution of reflexives in the way predicted by cue-based retrieval (Dillon et al., 2013).[1] Similar results have been found by Xiang et al. (2009), who showed that structural dependencies such as NPI licensing cannot be explained by a general cue-based retrieval mechanism.

A similar notion of accessibility is involved in resolution of other dependencies, too. In particular, it is known that quantificational distributive expressions like *no + noun phrase (NP)*, *every + NP*, *each + NP*, can only serve as antecedents for those pronouns that they c-command (Heim, 1982; Reinhart, 1983). In contrast to that, referential expressions like *the + NP*, *a + NP* can antecede pronouns that they do not c-command, because they support co-referential bindings – an option missing for quantificational distributive expressions. The contrast is shown in (4), from Heim (1982). (4-a) is grammatical under the interpretation that *he* refers back to *the soldier*, because no c-command is required to establish this relation. (4-b) is ungrammatical under the interpretation that *he* has *no soldier* as its antecedent, because the latter does not c-command the former (c-command does not span across clause boundaries).

(4) a. The soldier has a gun. Will he attack?
b. No soldier has a gun. *Will he attack?

The c-command restriction on pronoun resolution was investigated in several reading studies (e.g., Carminati et al., 2002; Cunnings et al., 2014, 2015; Koornneef et al., 2011; Moulton & Han, 2018). Here we discuss in more detail the study of Kush et al. (2015), which investigated the role of c-command from the perspective of cue-based retrieval. What they compared was a case in which the referential element either matched or mismatched the pronoun in gender ((5-a)–(5-b)) with a case in which the quantificational element matched or mismatched ((5-c)–(5-d)). Importantly, the referential and quantificational NPs in this example did not c-command the critical pronoun.

(5) a. **Referential, Match**
The troop leaders that *the girl scout* had no respect for had scolded *her* after the incident at scout camp.
b. **Referential, Mismatch**
The troop leaders that *the boy scout* had no respect for had scolded *her* after the incident at scout camp.

[1] More recently, however, it has been shown that Dillon et al. (2013), which tried to compare two dependency types within one experimental study, suffered from low power. Increasing the power led to results equivocal with cue-based retrieval (Jäger et al., 2020).

c. **Quantificational, Match**
The troop leaders that **no girl scout** had respect for had scolded *her* after the incident at scout camp.

d. **Quantificational, Mismatch**
The troop leaders that **no boy scout** had respect for had scolded *her* after the incident at scout camp.

They observed that in early and late measures, reading times were longer in the referential mismatch condition compared to referential match condition. Since in cases of referential mismatch, there was no antecedent present for the pronoun, the slowdown can straightforwardly be explained as indicating that readers did not know how to resolve the pronoun in (5-b) compared to (5-a).

Interestingly, in early measures, the effect was absent for quantificational elements, or even reversed. Matching quantificational phrases caused a (non-significant) slowdown in the reading of the pronoun and the subsequent text, compared to mismatching quantificational phrases. This is surprising from the perspective of cue-based retrieval and the three assumptions listed above. If establishing an anaphoric dependency between the pronoun and its antecedent happens in a way that is analogous to subject-verb dependencies, a speed-up due to matching would be expected, since the matching quantificational phrases should be retrieved faster than the mismatching one, even though neither of those can resolve the pronoun. Therefore, it seems that these dependencies are either established using a different mechanism than retrieval, or else it could be that structural relationships trump other features (also see Parker & Phillips, 2017).

In this work, we look at dependencies that are not unlike the dependencies used in Kush et al.'s design in that (i) they are anaphoric in nature and (ii) they are subject to constraints on accessibility. Our focus, however, will be on dependencies involving anaphoric pronouns that cross sentence boundaries. To illustrate how such dependencies are constrained, consider the contrast between (6) and (7), which shows that *a book* in (6-a) can antecede the pronoun *it* in (6-b), but it cannot do so in the case of (7).

(6) a. A man^1 read a book^2.
b. He$_1$ liked it$_2$.

(7) a. A man^1 didn't read a book^2.
b. #He$_1$ liked it$_2$.

Theories of meaning called dynamic semantics (see, e.g., Groenendijk & Stokhof, 1991; Heim, 1982; Kamp, 1981; Kamp & Reyle, 1993; Nouwen et al., 2016) aim to model the role of context on, among other things, pronoun resolution. In such theories, pieces of text, e.g., sentences, are seen as instructions to update context and the main focus is to understand what properties such an update has. We can already intuitively understand that the update that is performed by a comprehender after having read (6-a) is different from the update involved in reading (7-a). In dynamic semantics theories, this intuitive difference can be captured formally in terms of the type of update that is involved in each sentence, and this formal distinction can in turn be used to account for the observation that the pronoun *it* can refer back to *a book* in (6) but not in (7).

The dynamic semantic framework that we will use here to illustrate this is called Discourse Representation Theory (cf. Kamp 1981; Kamp & Reyle 1993; henceforth: DRT), and assumes that a hearer builds a mental model during comprehension. The mental model, called a discourse representation structure, is a representation that consists of two pieces of information: (i) discourse referents, which we can picture as pegs that serve as pointers to entities under discussion in that discourse, and (ii) discourse conditions, which specify what information the discourse provides on those entities. For example, assuming that (6-a) is the starting point of discourse, a comprehender would construct a mental model representation in (8) (see Brasoveanu & Dotlačil, 2020, for an explicit formalism for the incremental construction of mental representations).

(8) A man read a book. $\rightsquigarrow$

$$\begin{array}{|c|}\hline x\ y \\ \hline x \text{ is a man} \\ y \text{ is a book} \\ x \text{ read } y \\ \hline \end{array}$$

The top part of this representation shows the information that two entities were introduced in this short discourse, labeled (arbitrarily) as *x* and *y*. Furthermore, three pieces of information are collected on those entities, written as conditions on *x* and *y* in the bottom part. In this case, those conditions specify that the discourse referent labeled as *x* is a man, that the one labeled as *y* is a book, and that *x* read *y*.

When this sentence is followed by another one, as in (6), the current representation is further updated. Pronouns are interpreted just as old discourse referents, i.e., pegs that must have been introduced previously (e.g., *x* and *y* in this example). The resulting discourse representation structure is in (9), in which the contribution of the second sentence is captured in the last line of the discourse representation structure.

(9) He liked it. $\leadsto$

$$\begin{array}{|c|}\hline x\ y \\ \hline x \text{ is a man} \\ y \text{ is a book} \\ x \text{ read } y \\ x \text{ liked } y \\ \hline \end{array}$$

Crucially, in order to capture the fact that this latter update is not possible in the case of (7), DRT assumes, along with many other dynamic semantic frameworks, that it is the contribution of the negation that makes the peg introduced by *a book inaccessible* for a pronoun in a subsequent sentence. To implement this, DRT allows for embedding one discourse representation structure inside another. Intuitively, this can be understood as a situation in which inside one discourse, another (sub-)discourse is being developed. DRT precisely specifies under which conditions sub-discourses can be introduced (Kamp & Reyle, 1993). These conditions are not particularly relevant for us, and it suffices to note that negation and non-referential quantifiers are examples of triggers of sub-discourses. This means that the negated sentence in (7a) would be represented as in (10).

(10) A man did not read a book. $\leadsto$

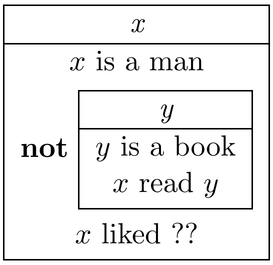
$$\begin{array}{|c|}\hline x \\ \hline x \text{ is a man} \\ \textbf{not}\ \begin{array}{|c|}\hline y \\ \hline y \text{ is a book} \\ x \text{ read } y \\ \hline \end{array} \\ \hline \end{array}$$

In this structure, the sub-discourse is preceded by **not**, signaling that the sub-discourse must not be true.

Importantly, discourse referents introduced in sub-discourses are not accessible outside of these sub-discourses. This condition restricts potential antecedents for pronouns. More concretely, it has a straightforward effect for (7-a). Since a new sentence updates the largest discourse representation structure, pronouns appearing in a new sentence cannot be interpreted as *y*, i.e., as a book. The only other discourse referent is *x*, so the only possible interpretation for pronouns is *x*, which could be paraphrased, given the conditions in (10), as ‘the man who read no book’.

Suppose that (10) is followed by a sentence like *he liked it*. This would lead to the update in (11). Here, the object of *liked* cannot be assigned a representation, since the only available discourse referent is *x* and *x* is a man, so it cannot be referred back to by *it*. Thus, it is predicted by DRT that the pronoun in this follow-up discourse will receive no interpretation, matching people’s intuitions (cf. Kamp & Reyle, 1993).

(11) He liked it. ⇝



In sum, Discourse Representation Theory is a semantic framework that provides (i) the interpretation of discourses and (ii) general conditions specifying whether a particular discourse is well-formed. With respect to the latter, DRT argues that pronouns can only be resolved to those discourse referents that are accessible. Like c-command, accessibility in DRT is a relational notion. However, unlike c-command, accessibility spans over sentences across the complete discourse. We can thus use accessibility to study how a condition on discourse well-formedness affects pronoun resolution in processing.

If we follow the traditional cue-based retrieval models and assume that the role of discourse accessibility in processing is similar to that of other features like gender or number, we would expect cue-based effects parallel to (2). In particular, when there is a target, a partially matching inaccessible distractor would slow down retrieval and, hence, processing. When there is no target, a partially matching inaccessible distractor would speed up retrieval and, hence, processing. If we follow DRT and assume that as far as pronoun resolution is concerned, accessibility is not just a feature but inaccessible discourse referents are simply ignored, we would expect that inaccessible discourse referents do not interfere with retrieval. If we find support for the DRT hypothesis, this could be implemented as some kind of pre-filter which is applied during processing before cue-based retrieval takes place, and which determines what elements should be considered by cue-based retrieval. Another possibility is assuming that accessibility is a very strong cue that overrides 'regular' cues.

3. Experiments 1a, 1b and 2

An eye-tracking reading experiment and two offline experiments were conducted to study cross-sentential anaphora and cue-based retrieval. We will first describe the items used in all these experiments.

The items for the experiments were all constructed in Dutch. In all the experiments presented here, every item took the form of a short narrative consisting of two sentences. In the first sentence, two discourse referents were introduced: the first one was introduced by a noun phrase that appears in the subject position, while the second one was introduced by a noun phrase in (prepositional) object position. The second sentence contained an anaphoric pronoun, whose interpretation depended on the discourse referents in the first sentence.

We made use of gender features to look into the effect of inaccessible antecedents on pronoun resolution. In particular, the gender of the noun phrases in the first sentence, and of the pronoun in the second sentence, were manipulated so that this pronoun either matched or mismatched the gender of the subject in the first sentence (S.MATCH vs. S.MIS). Similarly, the pronoun could either match or mismatch in gender with the object in the first sentence (O.MATCH vs. O.MIS). Orthogonal to these gender manipulations, the referentiality of the object (REF vs. NON-REF) was manipulated by varying the quantifier inside the object NP. In the REFerential conditions, the object was expressed by an indefinite NP containing the indefinite article *een* 'a', which made the object an accessible antecedent for the anaphoric pronoun in the subsequent sentence. In the NON-REFerential conditions, the object NP contained the non-referential quantifier *geen* 'no', which made the object inaccessible as an antecedent for the pronoun. An example of one item in all eight conditions is shown in (12) and (13), where gender of the relevant expressions is marked using subscripts for the original materials in Dutch; the English translation follows the original.

(12) **Referential conditions:**
a. De professor$_M$ heeft een zoon$_M$. De laatste jaren werkte hij$_M$...
The professor has a son. The last few years he worked... **s.match o.match**
b. De professor$_M$ heeft een dochter$_F$. De laatste jaren werkte hij$_M$...
The professor has a daughter. The last few years he worked... **s.match o.mis**
c. De professor$_M$ heeft een dochter$_F$. De laatste jaren werkte zij$_F$...
The professor has a daughter. The last few years she worked... **s.mis o.match**
d. De professor$_M$ heeft een zoon$_M$. De laatste jaren werkte zij*F*...
The professor has a son. The last few years she worked... **s.mis o.mis**
...helaas op alle feestdagen.
...unfortunately during all holidays.

(13) **Non-referential conditions:**
a. De professor$_M$ heeft geen zoon$_M$. De laatste jaren werkte hij$_M$...
The professor has no son. The last few years he worked... **s.match o.match**
b. De professor$_M$ heeft geen dochter$_F$. De laatste jaren werkte hij$_M$...
The professor has no daughter. The last few years he worked... **s.match o.mis**
c. De professor$_M$ heeft geen dochter$_F$. De laatste jaren werkte zij$_F$...
The professor has no daughter. The last few years she worked... **s.mis o.match**
d. De professor$_M$ heeft geen zoon$_M$. De laatste jaren werkte zij$_F$...
The professor has no son. The last few years she worked... **s.mis o.mis**
...helaas op alle feestdagen.
...unfortunately during all holidays.

To make sure that our gender manipulations were effective, kinship terms and other terms that are explicitly marked for gender (e.g. *vriend/vriendin* 'boyfriend'/'girlfriend') were used as the object of the first sentence. The pronoun in the second sentence was always explicitly marked for gender (*hij* ‘he’ vs. *zij* ‘she’). Because the singular 3rd person feminine pronoun *zij* (‘she’) is ambiguous with the plural 3rd person pronoun *zij* (‘they’), all of the second sentences involved subject-verb inversion, so that the singular marked verb always preceded the pronoun, thus making sure that the pronoun itself could not be misinterpreted as plural.

For the subject noun, only nouns that *stereotypically* referred to male characters were used, modeled on the design used in Sturt (2003). This was done so that it would be possible to create the S.MIS conditions without giving rise to uninterpretable discourses in the conditions where both the subject and the object mismatched the gender of the pronoun. In this way, despite a strong preference to interpret the subject as male, it would still be possible to reinterpret the character as female. This was important, because otherwise there would be no antecedent for the pronoun in a substantial amount of the experimental items, which might lead participants to give up on establishing an anaphoric relationships in general, or could lead them to employ alternative parsing strategies. The choice of the male-biased nouns was partly based on the nouns used by Sturt (2003) and mostly dependent on the fact that in Dutch, gender-biased nouns which can still be interpreted as the other gender are always male-biased. In other words, there are almost no nouns in Dutch which are stereotypically interpreted as female but can also refer to a male (e.g., a noun like *nurse* is explicitly marked for gender in Dutch). To make sure that the nouns we selected are indeed male-stereotypical, but can still refer to female referents as a last resort, we only included nouns that did not have a female-gender marked equivalent in Dutch (such as *loodgieter* ‘plumber’), or for which a female version was used in less than 0.7% of the time in the ‘Corpus Hedendaags Nederlands’ (http://corpushedendaagsnederlands.inl.nl). Additional evidence for the fact that these subject nouns indeed had a strong male bias was found in the resolution task discussed in the next subsection.

Recall that the items used here also aimed to exclude the possibility of accommodating the inaccessible antecedent as a potential antecedent for the pronoun. Concretely, to avoid the possibility of such accommodation, the materials adopted here not only made use of non-referential quantifiers to create non-referential object NPs, they also involved verbs of possession or creation (like *have* in (12) and (13)), in combination with the use of relational nouns like kinship terms as the object. This is relevant because, despite the fact that *no neighbors* in (14-a) is a non-referential NP and therefore forms an inaccessible antecedent for the pronoun in the subsequent sentence, readers can still accommodate the existence of the neighbors in context (i.e., the neighbors that the old man didn’t see), and use this as the antecedent for *they* in (14-b).

(14) a. The old man saw no neighbors in the last several days.
b. They always left for work early.

Such accommodation is impossible in (15), where the existence of a referent for the daughters of the old man in question is ruled out by the sentence in (15-a) itself. Consequently, the sentence in (15-b) is odd as a follow-up to (15-a) because there is no referent that can be accommodated as an antecedent for the pronoun *they*.

(15) a. The old man had no daughters.
b. #They always left for work early.

Thus, by making use of similar constructions in the items that are employed in the current studies, we rule out such accommodation, too. In Experiments 1a and 1b we show that, indeed, participants interpret the pronoun in the second sentence only very rarely to refer back to the non-referential NP in the first sentence.

3.1 Experiment 1a: Resolution task

3.1.1 Participants

Twenty-seven participants participated in an online study. The participants were found via social media. They indicated their native language and age. They could, but did not have to, indicate their gender. All participants self-identified as native speakers of Dutch. The mean age was 33.7 (SD: 14.6, range: 21–66). The majority of participants (21 in total) self-identified as female.

3.1.2 Design & procedure

To test whether the use of a non-referential quantifier indeed blocked the resolution of the object NP as an antecedent for the pronoun, a comprehension study was carried out.

For the purpose of the study, 32 items were created, following the design explained above, see (12) and (13) for an example. The items were presented online, along with a comprehension question which targeted the interpretation of the pronoun. For example, for (12)/(13), the question would ask: 'Who worked during holiday?' Participants could choose between three options: *professor* 'professor' (the subject), *zoon/dochter* 'son'/'daughter' (the object), or *anders* 'other'. The last response would be chosen when the participant thought another referent should resolve the pronoun or the participant was not sure which of the two (the subject or the object) should resolve the pronoun. Items were divided over 8 lists via a Latin Square and were randomly presented, interspersed with 32 fillers, which were of similar structures but did not include non-referential quantifiers. An example of a filler with the follow-up comprehension question is given here:

(16) a. De meeste mensen vonden achteraf toch dat de politicus gelijk had
The most people found afterwards still that the politician right had

gehad. Op het moment zelf was hij nog de zondebok geweest.
had. On the moment itself was he yet the scapegoat was.
'Most people realized afterwards that the politician had been right. Back then, however, he had been the scapegoat.'

b. Wie was er de zondebok geweest? politicus / mensen / anders
Who was there the scapegoat was? politician / people / other
'Who was the scapegoat? politician / people / other'

The experiment took around 15 minutes.

3.1.3 Results & discussion

Mean responses are summarized in **Table 1**.

Table 1: Mean responses per condition for the resolution task, Exp. 1a

Condition			**Response**		
Quantifier	**Subject**	**Object**	**Subject**	**Object**	**Other**
een (REF)	Match	Match	.59	.30	.11
	Match	Mismatch	.98	0	.02
	Mismatch	Match	.25	.66	.09
	Mismatch	Mismatch	.81	.14	.05
geen (NON-REF)	Match	Match	.94	.03	.03
	Match	Mismatch	.96	0	.04
	Mismatch	Match	.88	.07	.05
	Mismatch	Mismatch	.97	0	.03

We analyzed the data in the Bayesian paradigm using hierarchical models (Gelman et al., 2003; Gelman & Hill, 2006; McElreath, 2018; Nicenboim et al., 2021).

In Bayesian data analysis, one specifies the likelihood and the prior distributions over parameters of interest. The analysis results in posterior probability distributions of plausible values for a given model and data. We report medians and 95% credible intervals, i.e., the range of values for which we can be 95% certain that the true effect lies therein.

We used Bernoulli likelihood with logit link function. The dependent variable was the response with two values. For the purpose of the modeling, 'Object' and 'Other' responses were collapsed and treated as 0; 'Subject' responses were treated as 1. The fixed effects were SUBJECT (sum-contrast coded, match = 1, mismatch = –1), OBJECT (sum-contrast coded, match = 1, mismatch = –1), QUANTIFIER (sum-contrast coded, non-referential = 1, referential = –1) and all interactions. The model was fit with a full variance-covariance matrix, i.e., it was a so-called maximal model.

Following common practice, we use prior distributions that were "weakly informative" (Gelman et al., 2003, p. 55). This means that the priors contain little real-world knowledge and a priori they do not exclude any observable values. The prior distributions were constructed in a way which would not pull the results in any direction, since we wanted to remain agnostic about both the size and the direction of the effects. More concretely, the following prior distribution was assumed:

- Intercept: Normal($\mu = 0, \sigma = 3$)
- Fixed effects: Normal($\mu = 0, \sigma = 3$)
- Standard deviation of random effects: Normal($\mu = 0, \sigma = 3$), truncated at 0
- Random effects correlation: LKJ distribution with $\eta = 2$ (Lewandowski et al., 2009; Stan Development Team, 2021)

The model used 4 sampling chains, with 2,000 samples drawn from each chain. Half of these samples were discarded for warm-up; hence, the model had 4000 samples available for the analysis. All the parameters in the model had $\hat{R} \leq 1.05$, which supports model convergence.

The posterior distribution of the fixed effects is summarized in **Figure 2**. We focus on the effects that are clearly positive or negative, i.e., whose 95% credible intervals do not cross 0. The positive posterior distribution of SUBJECT shows that the subject matching in gender with the pronoun increases the preference for the subject-resolution of the pronoun. The negative posterior distribution of OBJECT shows that the object matching in gender with the pronoun decreases the preference for the subject-resolution of the pronoun. Both effects simply reveal that the match/mismatch of subject and object affect the pronoun resolution. More importantly for us, the positive effect of QUANTIFIER shows that non-referential quantifiers increase the preference for the subject-resolution of the pronoun. This means that the pronoun resolution is sensitive to the type of quantifier: the resolution of the pronoun towards the subject is stronger when the object quantifier is non-referential, or, in other words, non-referential objects are used less as pronoun antecedents.

There are also two interactions whose 95% credible intervals do not cross 0. The QUANTIFIER × SUBJECT interaction has a similar size but the opposite direction to the effect of SUBJECT. This interaction indicates that the effect of SUBJECT was modulated by quantifier type, in particular, non-referential quantifiers removed the effect of SUBJECT. In other words, when the object is non-referential, people resolve the pronoun to the subject even when the subject mismatches in gender. Such a resolution is possible, albeit pragmatically odd, and clearly preferred over considering a non-referential object as the antecedent. Second, there is a positive QUANTIFIER × OBJECT interaction, which shows that the role of match/mismatch of objects plays less of a role when the object is non-referential. Both interactions are in line with our claim that non-referential quantifiers were not used, or only minimally, for pronoun resolution.

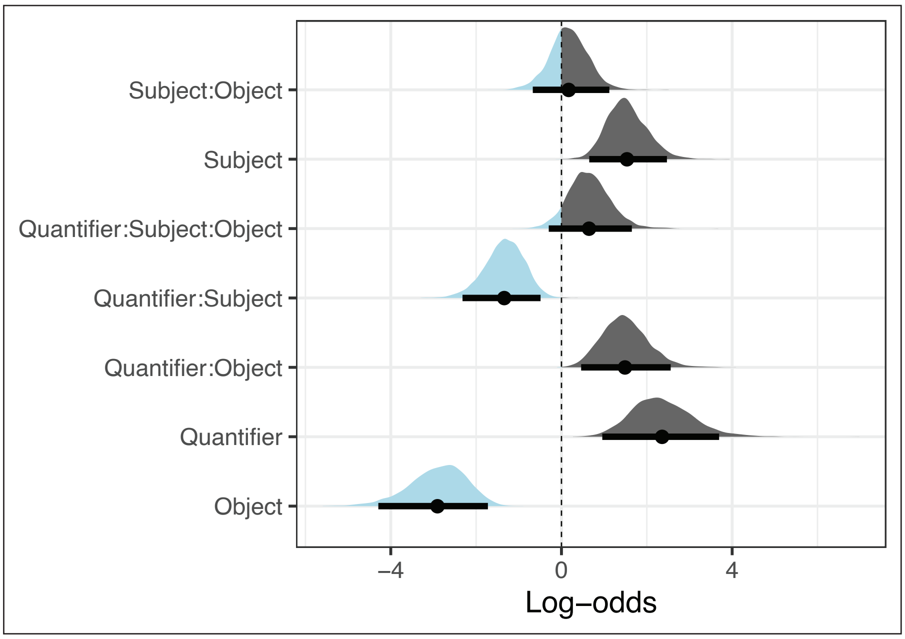


Figure 2: Effects in the preference task, Experiment 1a, given on the log-odds scale. The dot denotes the mean, the thick lines, the 95% credible intervals. Density areas higher than 0 appear in grey, density areas smaller than 0 appear in blue.

To further explore the difference between quantifier types on pronoun resolution, we consider a Bayesian model with nested comparisons, in which subject and object match/mismatch and their interactions are nested in the quantifier type. The random structure is maximal, the prior structure is identical to the previous model. The results are summarized in **Figure 3**. In nested comparisons, we see a clear positive effect of SUBJECT and a negative effect of OBJECT on pronoun resolution towards the subject for referential quantifiers. For non-referential quantifiers, the credible interval for SUBJECT spans positive and negative regions. The credible interval for OBJECT crosses 0, even though it is predominantly negative. This negative effect is observed for two reasons: (i) when we are close to the probability of 1, which is the case here, even small changes in preferences will result in large log-odds,[2] (ii) even when readers reject the subject interpretation of the pronoun, they select 'Other' as a response, rather than resolving the pronoun towards the object. In sum, interaction models and models with nested comparisons reveal that matching gender for subjects and objects affects pronoun resolution, but only in the case of referential quantifiers. Gender match/mismatch in non-referential quantifiers is almost completely ignored for the purpose of pronoun resolution: if there is any effect of such a match, it is very small (less than 1 percent decrease in the probability of selecting the subject, according to the Bayesian model with nested comparisons).

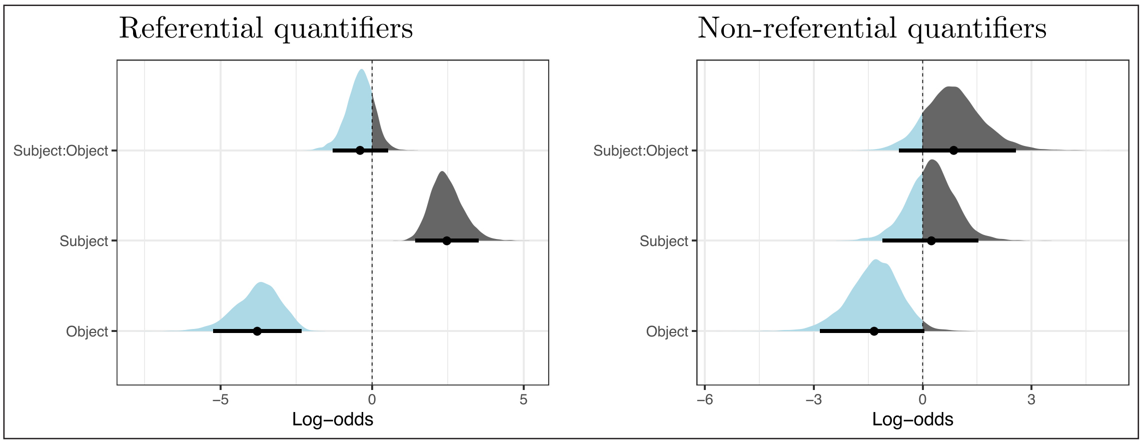


Figure 3: Comparisons nested in quantifier type in the preference task, Experiment 1a, given on the log-odds scale. The dot denotes the mean, the thick lines, the 95% credible intervals. Density areas higher than 0 appear in grey, density areas smaller than 0 appear in blue.

3.2 Experiment 1b: Acceptability judgement task

3.2.1 Participants

Sixty-eight participants (of whom 36 self-identified as female) were recruited via the online questionnaire platform Prolific. Their mean age was 27.21 years (SD: 8.07; range: 18–53). All participants reported being native speakers of Dutch; none of them reported suffering from dyslexia or other reading problems. The task lasted for 10–15 minutes, and participants were rewarded with 3 euros.

[2] Back-transforming the credible interval of OBJECT into probabilities shows the following interval: [–0.006, 0.00004].

3.2.2 Design & procedure

An online acceptability judgement task was carried out to test the acceptability of the stimuli in all conditions. The items were presented one by one, and participants were asked to judge how much sense each presented item made to them on a 7-point Likert scale (where 1 = completely uninterpretable, 7 = perfectly interpretable). The items were divided over 8 lists via a Latin Square and were randomly presented, interspersed with 48 fillers. Half of these fillers were contexts we expected to receive high scores, the other half we expected to receive low scores. Besides leading attention away from the goal of the experiment, the fillers served as an attention check. Visual inspection of the data led us to exclude 12 participants due to poor performance.

3.2.3 Results & discussion

Table 2 shows the descriptive statistics per condition.

Table 2: Mean and standard deviation of the scores per condition for the acceptability judgement task, Exp 1b.

Condition			**Response**	
Quantifier	**Subject**	**Object**	**Mean**	**SD**
een (REF)	match	match	4.374	1.832
	match	mismatch	4.074	1.822
	mismatch	match	4.853	1.779
	mismatch	mismatch	3.638	1.765
geen (NON-REF)	match	match	3.522	1.894
	match	mismatch	3.568	1.859
	mismatch	match	3.026	1.727
	mismatch	mismatch	3.312	1.899

Results were analyzed using Bayesian mixed-effects ordinal regression models (Bürkner & Vuorre, 2019). We used a cumulative model with a probit link function and so-called flexible thresholds (i.e., estimated distance between the different scores can vary). The dependent variable was the response (1–7). Fixed effects were sum-contrast coded in the same way as in Experiment 1a. The model was fit with a full variance-covariance matrix. The prior structure was as follows:

- Fixed effects: Normal($\mu = 0, \sigma = 5$)
- Standard deviation of random effects: Normal($\mu = 0, \sigma = 5$), truncated at 0
- Random effects correlation: LKJ distribution with $\eta = 2$ (Lewandowski et al., 2009; Stan Development Team, 2021)

The model used 4 sampling chains, with 6,000 samples drawn from each chain. Half of these samples were discarded for warm-up. All the parameters in the model had $\hat{R} \leq 1.05$, which supports model convergence.

The posterior distribution of the parameters is summarized in **Figure 4**. We see that matching subject and object increases acceptability, arguably because a match in gender ensures that the pronoun has an antecedent. Crucially, though, the positive credible interval of acceptability for OBJECT is modulated by the negative QUANTIFIER:OBJECT effect, which shows that the matching object affects acceptability mainly in referential objects. We also see a predominantly negative SUBJECT:OBJECT interaction, suggesting that mismatching subjects and objects decrease acceptability. This effect, however, is only present for referential quantifiers, as seen by the three-way positive interaction of QUANTIFIER × SUBJECT × OBJECT.

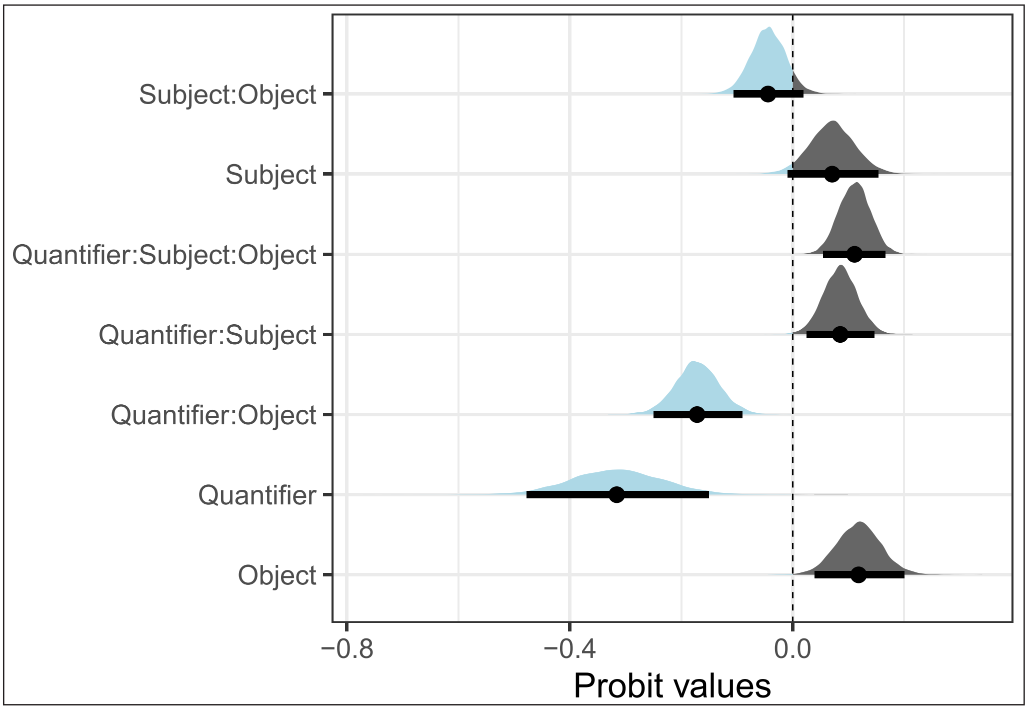


Figure 4: Effects in the acceptability task, Experiment 1b. The dot denotes the mean, the thick lines, the 95% credible intervals. Density areas higher than 0 appear in grey, density areas smaller than 0 appear in blue.

To further elucidate the difference in acceptability for referential and non-referential quantifiers, we also ran a model with nested comparisons, in which SUBJECT and OBJECT are nested in the referential/non-referential quantifier type. The random structure is maximal, the prior structure is identical to the previous model. The results, shown in **Figure 5**, reveal a clear contrast between referential and non-referential quantifiers.

In the case of the referential quantifier, mismatching objects decrease acceptability, arguably because their mismatch blocks one possible resolution for the pronoun, and when subjects and objects both mismatch, they also decrease acceptability, since no resolution is possible, unless one considers a non-stereotypical interpretation of the subject.

In the case of the non-referential quantifier, we only see a clear effect of SUBJECT, which, when mismatching, decreases acceptability. OBJECT clearly crosses 0; the interaction is positive but crosses 0 as well. In summary, mismatching subjects decrease acceptability, but objects do not. This is compatible with the claim that only subjects are considered for pronoun resolution in the case of non-referential quantifiers, since the gender of objects is irrelevant for the acceptability of the discourse.

After establishing, based on the acceptability and preference task, that non-referential objects in our design are not considered for pronoun resolution, or, at most, are considered for resolution in a very small number of cases, we turn to the eye-tracking study, which investigates the reading profile in pronoun resolution.

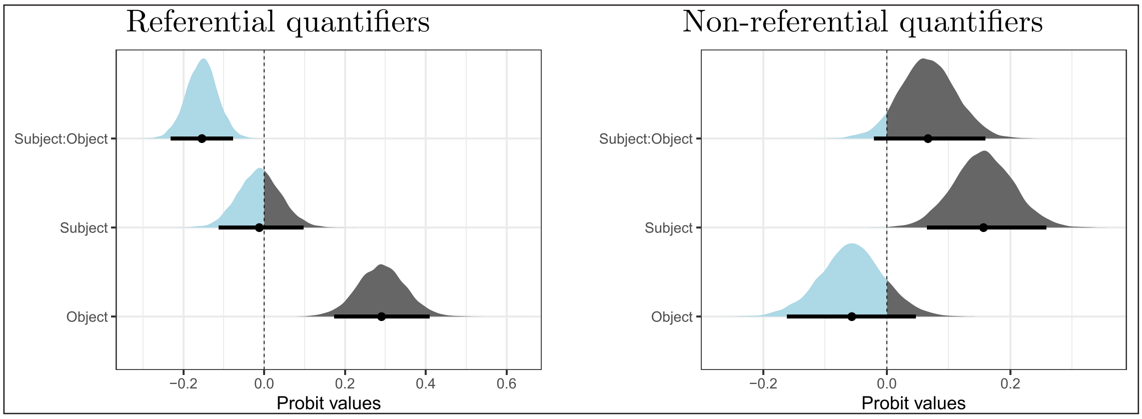


Figure 5: Comparisons nested in quantifier type in the acceptability task, Experiment 1b. The dot denotes the mean, the thick lines, the 95% credible intervals. Density areas higher than 0 appear in grey, density areas smaller than 0 appear in blue.

3.3 Experiment 2: Eye-tracking experiment

3.3.1 Participants

Forty-eight participants (41 female) participated in the experiment. Most of them were bachelor's or master's students at Utrecht University; all of them were acquired from the ILS participants database. The mean age was 23.37 years (SD: 3.01; range: 19–33). All participants were native speakers of Dutch. None of them reported suffering from dyslexia, severe eye abnormalities, or other reading problems. Participants with glasses or contact lenses were allowed to participate if their vision was corrected-to-normal. Participants were rewarded with 10 euros.

3.3.2 Materials and design

The experiment contained 32 target items and 48 fillers. The mean length of the target items was 102.4 characters; the mean length of the fillers was 109.1 characters. The target items all had the same structure, illustrated in **Table 3**.

Table 3: Example of the structure of a target item.

subject	**verb**	**object**	**introductory part**	**verb + pronoun**	**3 following words**	**wrap-up**
			pre-critical region	***critical region***	***post-critical region***	***wrap-up***
De professor	heeft	[een/geen] [zoon/ dochter].	De laatste paar jaar	werkte [hij/zij]	helaas op alle	feestdagen.
The professor	*has*	*[a/no] [son/ daughter].*	*The past few years*	*worked [he/she]*	*unfortunately on all*	*holidays.*

‘The professor has [a/no] [son/daughter]. The past few years, [he/she] unfortunately had to work during all the holidays.’

The fillers were short storylines, too, consisting of 2 or 3 sentences. Some fillers only contained a subject; some contained a subject and an object. In the second and/or third sentence, a reference to one of the characters was made, either by a pronoun or by a proper name/noun. Target items were divided over 8 lists via a Latin Square and, together with the fillers, presented in a unique random order for each participant. Maximally 2 items of the same condition could follow each other. Half of the fillers and half of the target items were followed by comprehension questions, which could be answered by yes or no. Each list started with a practice block. The practice block had a fixed order and consisted of four practice items, two of which were followed by a comprehension question (one to be answered with yes; one with no).

3.3.3 Procedure

The participants performed an eye-tracking-while-reading task in the ILS lab at Utrecht University. The experiment was programmed in ZEP (version 1.16) and the eye-tracking system used was Eyelink 1000, combined with a target sticker on the participant's forehead and a Beexy button box. Participants were seated in front of a PC monitor in a sound-proof, dimly lit booth. The distance to the screen was approximately 60 centimeters. After they read the study information and signed the consent form, the chair and camera were adjusted to the appropriate height/angle. Because of the COVID situation at that time, this all had to be done while taking social distance into account, which meant that it took more time than usual and that participants were sometimes asked to leave the booth to allow the experiment leader to go inside and make adjustments to the equipment.

Once the participants were seated in an appropriate position and read the instruction screen of the experiment, the first calibration and validation were performed, followed by the practice block. After a final opportunity to ask questions, a new calibration and validation were performed and the main experiment was started. In total, the experiment took 30–45 minutes.

The stimuli were horizontally aligned to the left side of the screen and consisted of multiple lines. Each line could maximally contain 70 characters. It was ensured that the critical region (the region with the pronoun) was always preceded by and followed by several words on the same line, making it appear roughly in the middle of the line. Before each stimulus was presented, a fixation trigger containing a drift correction was presented at the coordinates of the beginning of the stimuli. In this way, it was ensured that participants fixated on the position of the beginning of a stimulus before it was presented, and a new calibration could be performed if necessary.

Participants were instructed to press the lower (middle) button on the button box to proceed to the next stimulus. When a stimulus was followed by a comprehension question, they could answer it with 'no' by pressing the left button, and with 'yes' by pressing the right button. This information was visible at the bottom of the screen for all questions, to avoid confusion.

3.3.4 Data analysis

One participant was excluded from the analysis due to poor calibration; no participants were excluded based on the comprehension questions (all participants scored above 85% correct). The results were manually corrected for drift with the program Fixation by a student assistant, who did not know the purpose of the experiment and did not understand Dutch. Two items were excluded from the analysis due to typos that were discovered afterwards. In 6 participants, 1 or 2 items were excluded due to poor calibration or the absence of fixations (11 items in total). Data points without any fixations were excluded from the analysis (i.e., they were not treated as zeros).

Here we report models for two regions: the critical region and the post-critical region. The critical region (verb + pronoun) is the first region in which the pronoun can be resolved and in which the effect of the referentiality of the object can be measured. The post-critical region is the first spillover region and consists of the three words following the pronoun. Online we present results for the other regions as well.

The independent variables were SUBJECT (either matching or mismatching with the pronoun in gender), OBJECT (either matching or mismatching with the pronoun in gender), and QUANTIFIER (referential or non-referential). The variables had a sum-contrast coding: match was coded as 1, mismatch as –1 (for both SUBJECT and OBJECT); in the case of QUANTIFIER, the non-referential quantifier was coded as 1, the referential quantifier as –1. This means that a positive value for QUANTIFIER should be interpreted as showing that the condition with the non-referential quantifier took a longer time or increased regressions compared to the condition with the referential quantifier. A positive value for SUBJECT or OBJECT should be interpreted as showing that a condition with a matching subject or object was read longer or increased regressions compared to the condition with a mismatching subject or object.

The data were analyzed in the Bayesian paradigm using hierarchical models (Gelman et al., 2003; Gelman & Hill, 2006; McElreath, 2018; Nicenboim et al., 2021). As in Experiments 1a and 1b, the prior distributions used for modelling were "weakly informative" (Gelman et al., p. 55). Unless explicitly stated otherwise, models used 4 sampling chains, with 4000 samples drawn from each chain. Half of these samples were discarded for warm-up; hence, each model had 8000 samples available for the analysis. Trace plots were visually inspected to identify convergence issues. Additionally, only the models with all $\hat{R} \leq 1.01$, which suggests convergence, were used in the analyses.

Both models were fit with a full variance-covariance matrix. The models for the analysis of the eye-tracking data used a log-normal likelihood. We assumed a log-normal likelihood, since reading times data are approximately log-normally distributed, and therefore our model can resemble the data-generating process more closely (see Nicenboim et al., 2018; Rouder et al., 2008 for a more detailed discussion).

All the data analyses were conducted in the R software for statistical computing (R Core Team, 2021), and particularly with the use of the brms (Bürkner, 2017) and rstan (Stan Development Team, 2020) packages, which use Stan (Stan Development Team, 2021) probabilistic language.

The models took into account three predictors and their interactions as fixed effects: QUANTIFIER (REF vs. NON-REF), SUBJECT (MATCH vs. MIS), and OBJECT (MATCH vs. MIS). Participants and items were used as random effects.

We report the following measures:[3]

- Total Fixation Duration (TFD): Sum of all fixation durations in the coded region.
- Right-Bounded (RB): Sum of the durations of fixations that fall within the coded region before the region is left progressively for the first time.
- Probability of Regression (RP): Binary variable indicating whether there is a regression to a region with a lower code after the first pass (1) or not (0).

Two types of models were fit. The first type of model was used for the data in TFD and RB. It had a log-normal likelihood. For priors, it assumed a normal distribution with $\mu = 0$ and $\sigma = 10$ for the intercept, and a normal distribution with $\mu = 0$ and $\sigma = 1$ for slopes. For both the standard deviations of the random effects and the residual standard deviation, we used a truncated normal with $\mu = 0$ and $\sigma = 1$. For the random effects correlation between the intercept and the slope, we used the LKJ distribution (Lewandowski et al., 2009; Stan Development Team, 2021) with $\eta = 2$.

The second group of models was fit to RP. These models used Bernoulli likelihood with logit link function. They used the same predictors as the other models. The prior distribution of the intercept was narrower (a normal distribution with $\mu = 0$ and $\sigma = 1.5$). This is reasonable for models with logit link functions and was supported by prior predictive checks. These showed that normal distributions with wider σ resulted in most of the probability density being concentrated around 1 or 0, which is unreasonable. A priori, we should assume that the probability of regressing is concentrated around 0.5. The rest of the parameters used the same prior distributions as in the models described earlier.

3.4 Results

Before going into details about the results of the experiment, we want to very briefly recall the two different accounts discussed in Section 1, to highlight the expected effects. If we follow cue-based retrieval and assume that accessibility plays a role just as any other feature would, we would expect the following pattern for the non-referential (inaccessible) distractor (object): (i) when the subject matches, the matching non-referential object should lead to longer retrieval time and more erroneous retrieval compared to the mismatching non-referential object, which would translate into increased reading times and, arguably, regressions; (ii) when the subject mismatches, the matching non-referential object leads to shorter retrieval time compared to the mismatching non-referential object, which would translate into decreased reading times and,

[3] Other measures are presented in the online version of the paper: https://osf.io/xznw7/. The measures reported there are: first-pass reading times, re-reading (second-pass) times, and re-reading probability. These measures do not show any novel effects regarding the goal of the investigation, the role of accessible and inaccesible antecedents on reading in regions at or after pronoun resolution.

arguably, regressions. These predictions follow from the assumptions presented and exemplified in Section 2. If the object is accessible, then: (i) when the subject matches, the matching object has potentially no effect (since there is no erroneous retrieval) or might increase retrieval times, due to feature sharing; (ii) when the subject mismatches, the mismatching object slows down processing compared to the matching object (due to slower retrieval, but also due the fact that there is no potential antecedent for the pronoun).

If we assume, following DRT, that inaccessible elements are simply ignored for the purposes of resolution, e.g., because accessibility is some kind of pre-filter for cue-based retrieval, the predictions are different, but only for the inaccessible distractor. In that case, the match or mismatch with the inaccessible distractor should not affect reading times.

To translate this into effects in a statistical model, we note that the predictions of both accounts are compatible with a three-way interaction. However, the three-way interaction comes about in different ways. In particular, when we consider the inaccessible subcase, we should see no 2-way interaction effect between subject match/mismatch and object match/mismatch, if we stick to the predictions labeled as following DRT. In contrast, this 2-way interaction would be present under cue-based retrieval and the assumption that accessibility is a feature no different from gender/number for the purposes of retrieval. Because of this contrast, we report the results of the full model, but, in order to make the interactions easier to interpret, we also consider a model with nested comparisons. In the latter model, SUBJECT and OBJECT are nested in referential and non-referential QUANTIFIER, and we are particularly interested in SUBJECT and OBJECT effects in NON-REFERENTIAL QUANTIFIER.

In the tables, we report a mean of the estimate, followed by the lower and the upper bound of 95% credible intervals. The detailed results of all the measurements are available online.[4]

3.4.1 Critical region: verb + pronoun

The critical region consists of the verb and the pronoun. It is the first region in which the pronoun can be resolved and in which an effect of inaccessibility of the object can be measured. Descriptive summaries of the measures presented in this section are shown in **Table 4**. Details of the effects for this region are shown in **Table 5**.

On right-bounded reading times (RB), we observed a negative effect of SUBJECT. This means that in the mismatching subject condition, the participants spent more time in that region before leaving to the right than in the matching subject condition. In the case of OBJECT, we also observed a predominantly negative effect, which, however, spans 0.

[4] https://osf.io/xznw7/.

Table 4: Critical region: Mean raw reading times/count summaries by measure and condition.

Quantifier	**Subject**	**Object**	**RB (ms)**		**TFD (ms)**		**RP (pct)**
			mean	**SE**	**mean**	**SE**	
een (REF)	match	match	271.20	11.44	411.77	21.94	7.74
	match	mis	297.64	14.52	442.29	22.14	13.04
	mis	match	308.12	16.29	440.06	22.41	11.26
	mis	mis	310.69	13.15	541.26	26.31	13.46
geen (NON-REF)	match	match	288.36	14.12	415.93	22.58	15.75
	match	mis	289.42	10.91	444.88	23.30	7.14
	mis	match	304.82	12.30	547.32	26.97	14.29
	mis	mis	331.58	17.45	543.96	27.30	19.18

Table 5: Critical region: Summary of the results. Effects with credible intervals that do not cross 0 are boldfaced.

Measure	**Variable**	**Estimate**	**Q2.5**	**Q97.5**
RB (log-ms)	**Subject**	**–0.039**	**–0.066**	**–0.011**
	Object	–0.025	–0.053	0.002
	Quantifier	0.013	–0.013	0.04
	Subject × Quantifier	–0.008	–0.037	0.021
	Object × Quantifier	0.001	–0.025	0.027
	Subject × Object	0.006	–0.034	0.023
	Subject × Object × Quantifier	0.001	–0.031	0.035
TFD (log-ms)	**Subject**	**–0.087**	**–0.123**	**–0.050**
	Object	**–0.045**	**–0.078**	**–0.011**
	Quantifier	0.029	–0.006	0.064

(Contd.)

Measure	Variable	Estimate	Q2.5	Q97.5
	SUBJECT × QUANTIFIER	**–0.031**	**–0.061**	**–0.001**
	OBJECT × QUANTIFIER	0.029	–0.004	0.063
	SUBJECT × OBJECT	0.007	–0.023	0.038
	SUBJECT × OBJECT × QUANTIFIER	–0.028	–0.063	0.006
RP (log-odds)	SUBJECT	–0.222	–0.509	0.053
	OBJECT	–0.058	–0.317	0.186
	QUANTIFIER	0.137	–0.157	0.444
	SUBJECT × QUANTIFIER	–0.074	–0.308	0.16
	OBJECT × QUANTIFIER	0.19	–0.046	0.44
	SUBJECT × OBJECT	0.15	–0.084	0.402
	SUBJECT × OBJECT × QUANTIFIER	0.196	–0.086	0.466

Somewhat similar results were observed on total fixation duration (TFD). The effect of SUBJECT was negative, the effect of OBJECT was also negative, more clearly than in the case of RB, but less so compared to SUBJECT. Finally, the effect of QUANTIFIER was predominantly positive. An important effect is the interaction between the subject or object and the quantifier, as it can give us direct insight into the influence of the inaccessible antecedent. On TFD, we observed a negative interaction effect between SUBJECT and QUANTIFIER, and a comparably large positive interaction of OBJECT and QUANTIFIER. A graphical summary of RB and TFD posterior distributions is shown in **Figure 6**.

We explore the interactions in more detail in the nested model of TFD; see **Table 6**. We start with the SUBJECT × QUANTIFIER interaction. Inspecting this model for TFD reveals that under both conditions, a mismatching subject evoked longer reading times. The mean of the effect was further from 0 for the non-referential quantifier than for the referential quantifier. Therefore, for both quantifiers, a mismatching subject resulted in a slowdown, but in the case of the non-referential quantifier, the slowdown was larger. A similar analysis of the OBJECT × QUANTIFIER interaction shows that in the case of the referential quantifier, the effect of OBJECT was clearly negative, while in the case of the non-referential quantifier, there was no such clear effect (the credible interval crosses 0). Hence, there was some slow-down on the mismatching object in the former case, but none or almost none in the latter case. The SUBJECT × OBJECT interactions cross 0 in referential as well as non-referential quantifier conditions.

Table 6: Critical region: Summary of the effects in the model with comparisons nested by quantifier. Effects with credible intervals that do not cross 0 are boldfaced.

Measure	**Quantifier**	**Variable**	**Estimate**	**Q2.5**	**Q97.5**
RB (log-ms)	REF	SUBJECT	–0.031	–0.074	0.012
		OBJECT	–0.025	–0.060	0.010
		SUBJECT × OBJECT	–0.007	–0.053	0.041
	NON-REF	**SUBJECT**	**–0.046**	**–0.083**	**–0.009**
		OBJECT	–0.024	–0.064	0.014
		SUBJECT × OBJECT	–0.004	–0.043	0.034
TFD (log-ms)	REF	**SUBJECT**	**–0.055**	**–0.105**	**–0.006**
		OBJECT	**–0.073**	**–0.122**	**–0.023**
		SUBJECT × OBJECT	0.036	–0.011	0.083
	NON-REF	**SUBJECT**	**–0.118**	**–0.161**	**–0.073**
		OBJECT	–0.017	–0.064	0.03
		SUBJECT × OBJECT	–0.021	–0.064	0.023
RP (log-odds)	REF	SUBJECT	–0.126	–0.445	0.2
		OBJECT	–0.215	–0.519	0.07
		SUBJECT × OBJECT	–0.062	–0.369	0.255
	NON-REF	SUBJECT	–0.298	–0.630	0.026
		OBJECT	0.145	–0.231	0.518
		SUBJECT × OBJECT	**0.348**	**0.014**	**0.684**

The probability of regression (RP) showed a predominantly negative posterior distribution for SUBJECT in **Table 5**. QUANTIFIER and OBJECT showed no effect that was clearly negative or positive. We thus see that mismatching subjects increased the chance of regressions, which corresponds to the observed increased reading times on RB and TFD, due to mismatching subjects. Examining the nested model shows no clear effect of SUBJECT or OBJECT. There is an interesting effect of SUBJECT × OBJECT interaction in the non-referential condition. This interaction is in line

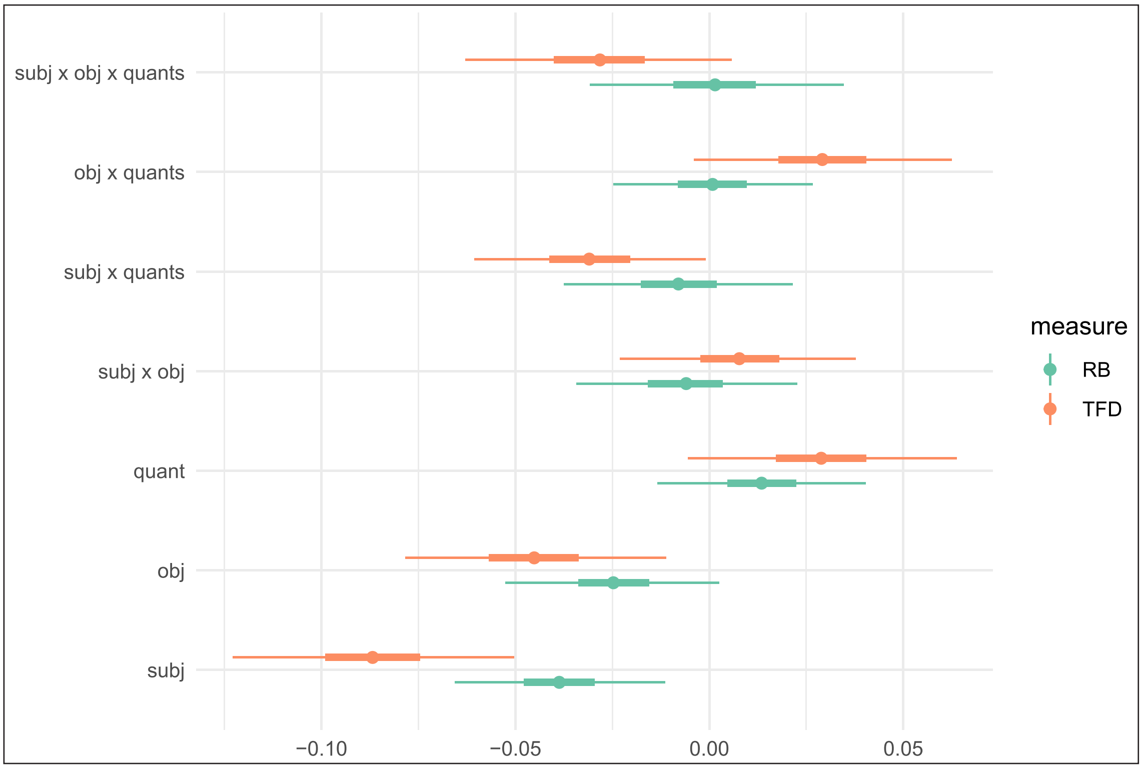


Figure 6: Effects measured on TFD and RB in the critical region in log-ms. The dot represents the mean, the thick lines, 50% credible intervals. The thin lines represent 95% credible intervals.

with the predictions of the cue-based retrieval model; however, somewhat unexpectedly from that perspective, the interaction effect is accompanied by a slightly positive posterior distribution of OBJECT, which suggests that people generally regress more when the object matches.

The emerging picture is that we see a robust effect of SUBJECT: mismatching subjects slow down reading times in both early and late reading measures and increase the probability of regression in the critical region. The effect of a mismatching object is less clear-cut. OBJECT is predominantly negative in RB, even though the 95% credible interval crosses zero in that case. In TFD, it is accompanied by a positive OBJECT × QUANTIFIER interaction, which shows that the mismatching object only causes observable difficulties when it is combined with a referential quantifier. The difficulties of object mismatch are diminished or completely disappear when the quantifier is non-referential, which is supported by the fact that when we consider the nested model, it reveals the negative effect of OBJECT only for referential quantifiers. RP is the only measure which reveals an interaction of SUBJECT × OBJECT in the non-referential condition. This interaction, we noted, would be in line with the predictions of cue-based retrieval.

3.4.2 Post-critical region (3 words following the pronoun)

The post-critical region consisted of the three words following the critical region. Descriptive summaries are provided in **Table 7**; details of the models of the results in this region are provided in **Table 8**.

Table 7: Post-critical region: Mean raw reading times/count summaries by measure and condition.

Quantifier	**Subject**	**Object**	**RB (ms)**		**TFD (ms)**		**RP (pct)**
			mean	**SE**	**mean**	**SE**	
Ref	match	match	549.77	23.58	851.79	41.52	24.86
	match	mis	558.49	32.49	809.33	36.90	25.71
	mis	match	576.07	28.71	800.90	37.49	31.58
	mis	mis	637.55	34.54	964.53	52.95	40.23
Non-ref	match	match	537.67	26.08	797.75	33.63	29.82
	match	mis	528.80	22.61	762.00	34.79	28.74
	mis	match	612.00	34.65	955.45	46.76	32.18
	mis	mis	605.37	32.79	893.17	44.48	35.09

Table 8: Post-critical region: Summary of the results. Effects with credible intervals that do not cross 0 are boldfaced.

Measure	**Variable**	**Estimate**	**Q2.5**	**Q97.5**
RB (log-ms)	**Subject**	**–0.046**	**–0.082**	**–0.011**
	Object	–0.008	–0.043	0.027
	Quantifier	–0.005	–0.034	0.023
	Subject × Quantifier	0.002	–0.03	0.034
	Object × Quantifier	0.001	–0.029	0.028
	Subject × Object	0.022	–0.006	0.052
	Subject × Object × Quantifier	–0.018	–0.048	0.013

(Contd.)

Measure	Variable	Estimate	Q2.5	Q97.5
TFD (log-ms)	**Subject**	**–0.044**	**–0.074**	**–0.015**
	Object	0.003	–0.029	0.035
	Quantifier	< 0.001	–0.03	0.031
	Subject × Quantifier	–0.028	–0.059	0.002
	Object × Quantifier	0.03	0.002	0.057
	Subject × Object	0.025	–0.002	0.052
	Subject × Object × Quantifier	**–0.029**	**–0.057**	**–0.001**
RP (log-odds)	**Subject**	**–0.275**	**–0.455**	**–0.098**
	Object	–0.084	–0.236	0.069
	Quantifier	0.031	–0.144	0.198
	Subject × Quantifier	0.106	–0.1	0.311
	Object × Quantifier	0.058	–0.116	0.235
	Subject × Object	0.094	–0.075	0.255
	Subject × Object × Quantifier	–0.027	–0.185	0.129

In the RB measure, a slowdown due to subject mismatch was found. Furthermore, two interactions were predominantly positive/negative. First, the interaction of subject × object was positive. Second, the interaction of subject × object × quantifier was negative. The interactions should be interpreted as follows. The first interaction reveals that the subject mismatch combined with the object mismatch led to an additional slowdown. The second interaction reveals that the slowdown due to the subject and the object mismatch was different for the referential and the non-referential quantifier.

We can make the interpretation clearer when we consider the nested model. Data of the models are in **Table 9**. The subset with the referential quantifier reveals a negative effect of subject. It also reveals a predominantly positive effect of subject × object. In other words, when reading sentences with the referential quantifier, readers slowed down when the subject mismatched the pronoun, and they also slowed down when both subject and object mismatched the pronoun. The subset with the non-referential quantifier also shows a negative effect of subject. However, the interaction effect is close to zero, i.e., the posterior distribution is not predominantly positive or negative. This shows that in sentences with the non-referential quantifier, readers slowed down when the subject mismatched the pronoun, while the combination of subject and object mismatch did not slow down readers any further.

Table 9: Post-critical region: Summary of the effects in the model with comparisons nested by quantifier. Effects with credible intervals that do not cross 0 are boldfaced.

Measure	Quantifier	Variable	Estimate	Q2.5	Q97.5
RB (log-ms)	Ref	Subject	–0.048	–0.103	0.005
		Object	–0.007	–0.048	0.035
		Subject × Object	0.04	–0.002	0.081
	Non-ref	**Subject**	**–0.044**	**–0.088**	**–0.001**
		Object	–0.009	–0.052	0.035
		Subject × Object	0.004	–0.039	0.047
TFD (log-ms)	Ref	Subject	–0.016	–0.056	0.024
		Object	–0.027	–0.071	0.015
		Subject × Object	**0.054**	**0.014**	**0.093**
	Non-ref	**Subject**	**–0.073**	**–0.118**	**–0.028**
		Object	0.033	–0.008	0.073
		Subject × Object	–0.003	–0.042	0.035
RP (log-odds)	Ref	**Subject**	**–0.388**	**–0.703**	**–0.092**
		Object	–0.146	–0.373	0.078
		Subject × Object	0.117	–0.105	0.344
	Non-ref	Subject	–0.160	–0.385	0.066
		Object	–0.029	–0.28	0.208
		Subject × Object	0.065	–0.167	0.3

A situation similar to what we described for RB is also seen for TFD. In **Table 8**, the total fixation duration reveals a negative effect of subject. There is also a predominantly positive interaction for subject × object. This positive interaction is modulated by the negative three-way interaction of SUBJECT × OBJECT × QUANTIFIER. Focusing on the SUBJECT × OBJECT interactions, we see that the combination of the subject and object mismatch shows a different pattern in referential and non-referential quantifiers. **Table 9** shows that the interaction SUBJECT × OBJECT is clearly positive for the referential case, but not for the non-referential quantifier. This is in line with the DRT predictions we mentioned above. We also see in **Table 9** that the posterior distribution of OBJECT leans negative and spreads out across zero for referential quantifiers, while it leans positive for non-referential quantifiers. Focusing on the latter case, the positive effect suggests that *matching* non-referential quantifiers lead to higher reading times. This is opposite to the effect for OBJECT that we observed so far. However, it fits well with the findings from the acceptability study, which revealed a numerical trend towards lower acceptability of matching non-referential quantifiers. The same explanation that was proposed there might explain the effect here. The posterior distributions of parameters in RB and TFD are also summarized in **Figure 7**.

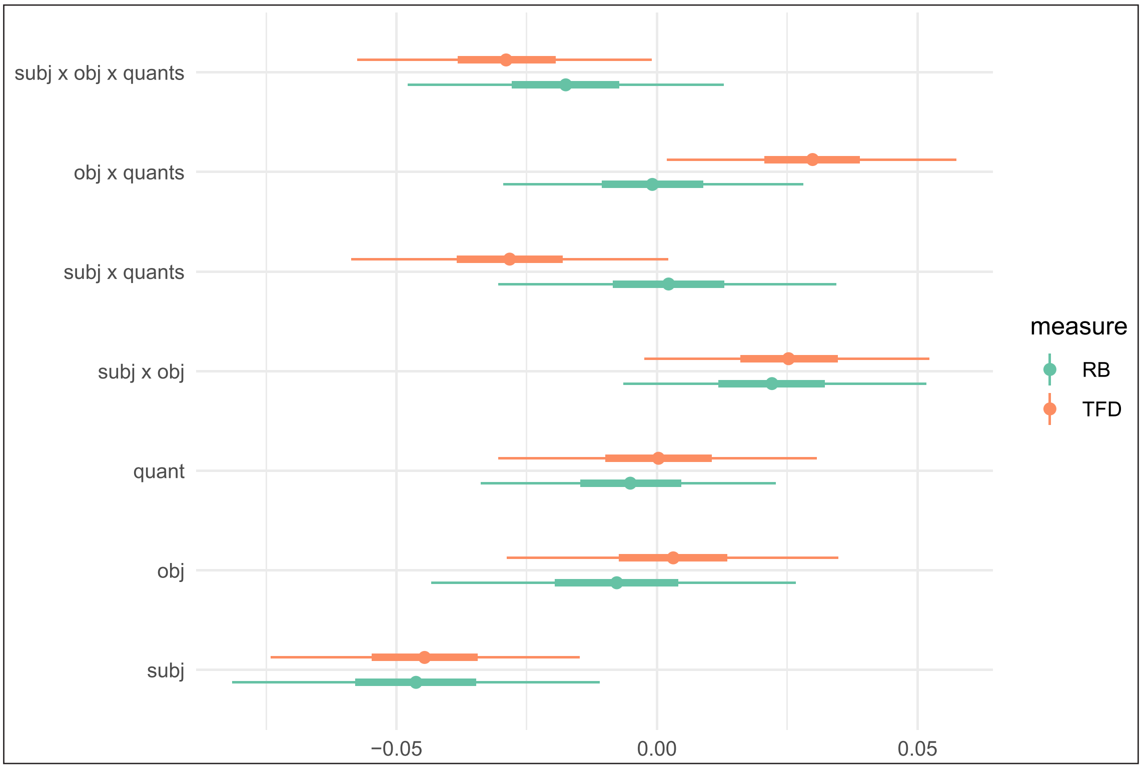


Figure 7: Effects measured on TFD, RB, and RRD at the post-critical region in log-ms. The dot represents the mean, the thick lines, 50% credible intervals. The thin lines represent 95% credible intervals.

In RP, the only effect that was clearly strongly positive/negative was the effect of SUBJECT, i.e., mismatching subjects increased regressions from the region. The nested model shows that the effect of SUBJECT is negative for both the referential and the non-referential case, even though the former negative effect is more pronounced.

In sum, we see that the post-critical region shows a very consistent effect of mismatching subject, which causes a slowdown in early and late reading measures and increases regressions from the region. Aside from the effect of mismatching subject, we also observe the combined effect of mismatching subjects and objects in RB and TFD. This interaction, however, is modulated by the three-way interaction of SUBJECT × OBJECT × QUANTIFIER (stronger in the case of TFD). Nested models show that the slowdown due to mismatching subjects and objects is driven by sentences in which the quantifier is referential, while the sentences in which the quantifier is non-referential do not show this effect. This strongly suggests that the mismatching object affects reading in the post-critical region, but only when it is accessible, i.e., when it appears as a referential quantifier.

3.5 Discussion

A first important observation in the results in 3.4 is that we found a clear slowdown when the stereotypical gender of the subject in the preceding sentence mismatched the gender of the pronoun. This was found in the critical and post-critical region for all measures shown (RB, TFD, and RP). The typicality effect observed in previous experiments (e.g., Sturt, 2003) was thus clearly replicated. This is important to note, as it means that our experimental manipulation was effective, and the match and mismatch of the gender cue we used were reacted to as expected.

Second, the experiment showed an effect of mismatching object in total fixation duration and regression probability. Regarding total fixation duration, the crucial observation is that the object effect interacted with referentiality. This was visible in the nested models, which showed that object mismatch (in the critical region) and subject-object interaction (in the post-critical region) were clearly visible only in the referential condition. The only case in which the credible interval of the subject-object-quantifier three-way interaction excluded 0, which was uncovered in total fixation duration in the post-critical region, goes in line with DRT predictions: the three-way interaction comes about due to the subject-object interaction in the referential condition and the absence of this interaction in the non-referential condition. This is compatible with the interpretation that the object gender match or mismatch only plays a role when the object is referential.

Do we see any interference driven by the inaccessible (non-referential) element that would follow the pattern predicted by cue-based retrieval? If so, we would expect a positive effect of OBJECT in the subset of the data that only includes the non-referential quantifier and the matching subject (since the matching object could act as a partially matching distractor and either increase a fan or increase the chance of being erroneously recalled; see the assumptions in Section 2). Furthermore, we would expect a negative effect of OBJECT in the subset of the data that only includes the non-referential quantifier and *mismatching* subject (since the match in object increases the chance of recall that is (erroneously) accepted when the subject is hard to accept as the antecedent of the pronoun; see the assumptions in Section 2).

The pattern predicted by cue-based retrieval is observed in one case, namely, in regression probability in the critical region, which shows a subject-object interaction in the non-referential condition at the critical region. There are three challenges to this interpretation, though. First, the credible interval crosses 0 in the case of the subject-object-quantifier three-way interaction, suggesting that whatever effect we observe in the non-referential condition cannot be fully distinguished from the referential condition. Crucially, the predictions of cue-based retrieval that we summarized above hold for the non-referential condition, not the referential case. Second, the descriptive summary in **Table 4**, suggests that the interaction is driven by the effect of matching/mismatching object on regressions when the accessible antecedent, i.e., the subject, matches the pronoun. However, the cue-based retrieval model of agreement and anaphora was mainly supported from data in which the accessible antecedent *mismatched* the resolution element (e.g., Jäger et al., 2017). Finally, it would be good if this pattern was observed in other measures beyond regressions, since the cue-based retrieval model is usually supported not (just) in regressions, but also in reading time measures, and the original cue-based retrieval model of Lewis & Vasishth (2005) was developed to predict reading times rather than regressions.[5]

To investigate the last issue, we zoom in on the effect of mismatching object on reading time measures. Furthermore, we check only those cases in which the subject mismatched the pronoun. We summarize the object effects in **Figure 8** for the critical and post-critical region and for RB and TFD. We would expect a negative effect. There is a very weak tendency of RB to go in the negative direction, but in general, the posterior distributions are almost symmetrically spread around zero, suggesting no effect of object match/mismatch. This finding in these reading measures is in line with the assumption that inaccessible antecedents are ignored.

4. Follow-up experiment with larger statistical power

A potential problem with Experiment 2 is a lack of statistical power. It might be possible that there actually is an effect of the object, but that the experiment was not powerful enough to find it. Therefore, we decided to perform a follow-up experiment with increased power. In this new experiment, we only used the non-referential condition, which reduced the number of conditions from 8 to 4. We also tested 56 participants, rather than 48. Furthermore, we increased the number of target items from 32 to 40. With some minor adjustments to some of the stimuli, we attempted

[5] But see Dotlačil (2021) and Rabe et al. (2023) on how regressions could be modeled when cue-based retrieval is combined with the E-Z reader and SWIFT model, respectively.

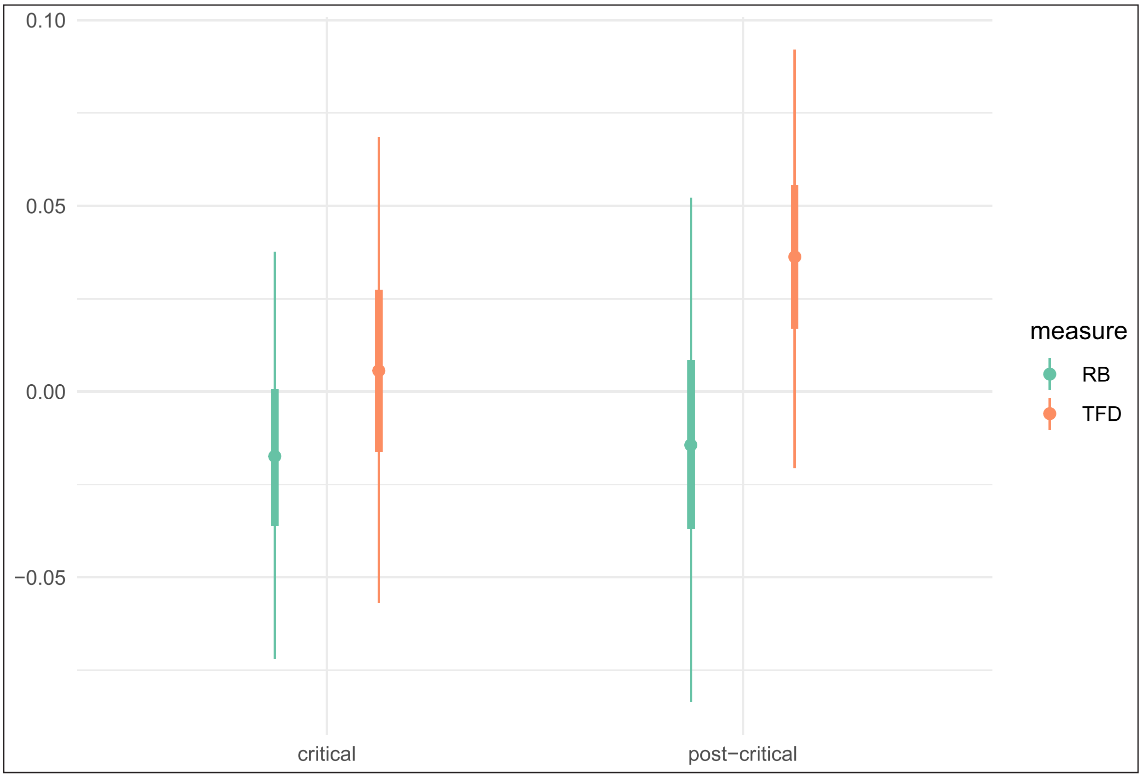


Figure 8: Effects of object match measured on the subject-mismatch subset of TFD, RB in log-ms and two regions of interest. The dot represents the mean, the thick lines, 50% credible intervals. The thin lines represent 95% credible intervals.

to even further reduce the possibility of accommodation of the inaccessible antecedent. Because of these minor adjustments and the extra stimuli, we also repeated the resolution task (Experiment 3a) and the acceptability judgement task (Experiment 3b). The results will be summarized below, before we turn to the eye-tracking experiment.

4.1 Experiment 3a: Resolution task

We tested 28 participants, of whom 22 self-identified as female. All of them were native speakers of Dutch; they were recruited via the ILS participant database. They were mostly students from Utrecht University, and their mean age was 20.9 (SD: 2.90; range: 18–28). The task was performed after they participated in the follow-up eye-tracking study.

The design of the experiment followed the design of Experiment 1a. We wanted to see how often participants select the subject and the object as the resolution of the pronoun. When looking at the individual items, it appears that there are two items in which the object was selected as the referent relatively often when it matched the pronoun. Both of these items do allow some accommodation and, therefore, we decided to exclude them. After excluding these two items, we see that the object was only selected as the antecedent in around 2% of cases when it matched the pronoun, and in 0.5% of cases when it mismatched. These are even lower numbers than in Experiment 1a, where the negative object was selected in around 7% of the cases when the object matched and the subject mismatched. This means that the possibility that the non-referential object might be used for pronoun resolution is extremely low in our items, and arguably impossible. The results are summarized in **Table 10**; see **Table 1** for the comparison.

Table 10: Mean responses per condition for the resolution task.

Condition		**Response**		
Subject	**Object**	**Subject**	**Object**	**Other**
Match	Match	0.95	0.02	0.03
Match	Mismatch	0.99	0.01	0
Mismatch	Match	0.95	0.02	0.03
Mismatch	Mismatch	0.97	0	0.03

We analyzed the data using a Bayesian mixed-effects model. The model used Bernoulli likelihood with logit link function with the dependent variable the response ('Object' and 'Other' responses were collapsed and treated as 0, 'Subject' responses were treated as 1) and fixed effects SUBJECT (sum-contrast coded, match = 1, mismatch = –1), OBJECT (sum-contrast coded, match = 1, mismatch = –1) and their interactions. The same model (but with the additional factor of QUANTIFIER) was used in Experiment 1a, presented in 3.1. See that section for more details on the model.

The posterior distribution of the fixed effects is summarized in **Figure 9**. We match the results in the previous preference task, Experiment 1a. We observe no clear effect of SUBJECT. Furthermore, the credible interval of OBJECT is predominantly negative, even though it does cross 0 (for a direct comparison, see the right graph of **Figure 3**). The negative effect of Object is due to two facts: (i) when we are close to the probability of 1, as is the case here, even small changes in preferences will result in large log-odds,[6] (ii) even when readers reject the subject interpretation of the pronoun, they select 'Other' as a response, rather than resolving the pronoun towards the object.

[6] Back-transforming the credible interval of OBJECT into probabilities shows the following interval: [–0.02, 0.0002].

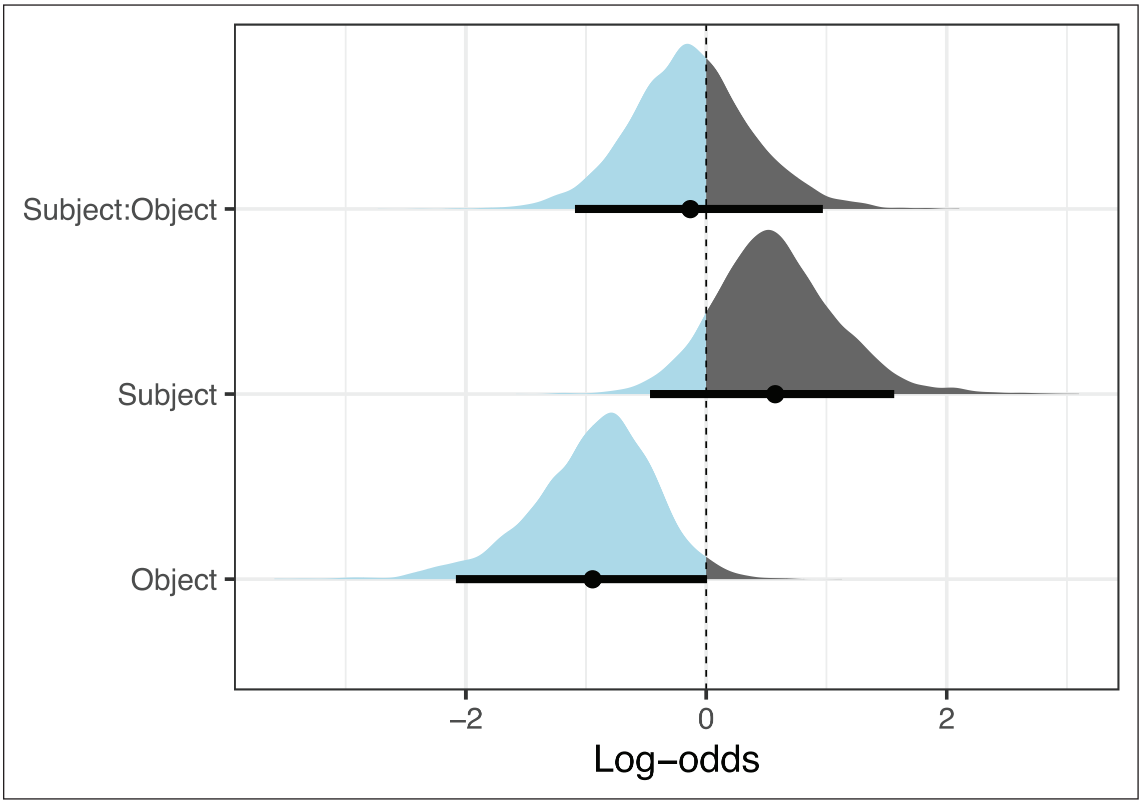


Figure 9: Effects in the preference task, Experiment 3a, given on the log-odds scale. The dot denotes the mean, the thick lines, the 95% credible intervals. Density areas higher than 0 appear in grey, density areas smaller than 0 appear in blue.

4.2 Experiment 3b: Follow-up acceptability judgement task

We tested 28 participants, of whom 20 self-identified as female. All of them were native speakers of Dutch; they were recruited via the ILS participant database. They were mostly students from Utrecht University, and their mean age was 23.8 (SD: 3.92; range: 18–32). The task was performed after they took part in the eye-tracking study. The design of the experiment followed Experiment 1b. Four participants were excluded due to poor performance. The two items that were detected in the pronoun resolution task as easily allowing accommodation, which were consequently excluded in eye-tracking, were also excluded in the acceptability judgement task.

Table 11 shows the descriptive statistics per condition. We expected that conditions with the matching subject would receive a higher rating than conditions where the subject mismatched since the subject could always resolve the pronoun, but in case of a mismatching gender this required additional reasoning. As the object was always non-referential in this experiment, it could never resolve the pronoun. Therefore, we expected that an effect of object mismatch would be smaller or absent.

Table 11: Mean and standard deviation of the scores per condition.

Condition		Response	
SUBJECT	OBJECT	Mean	SD
match	match	4.64	1.82
match	mismatch	4.65	1.87
mismatch	match	4.15	1.90
mismatch	mismatch	4.32	1.89

The results reflect these expectations. They were analyzed using Bayesian mixed-effects ordinal regression models with a probit link function and flexible thresholds, just as in Experiment 1b presented in 3.2. See that section for more details on the model. The posterior distribution of the parameters is summarized in **Figure 10**. The effect of SUBJECT, as expected, is positive, showing that mismatching subjects decrease acceptability, even though the 95% credible interval crosses 0. Neither OBJECT nor the SUBJECT × OBJECT interaction are clearly positive/negative, which shows that there was no clear effect of object match/mismatch on acceptability, arguably because objects were ignored for the purposes of pronoun resolution.

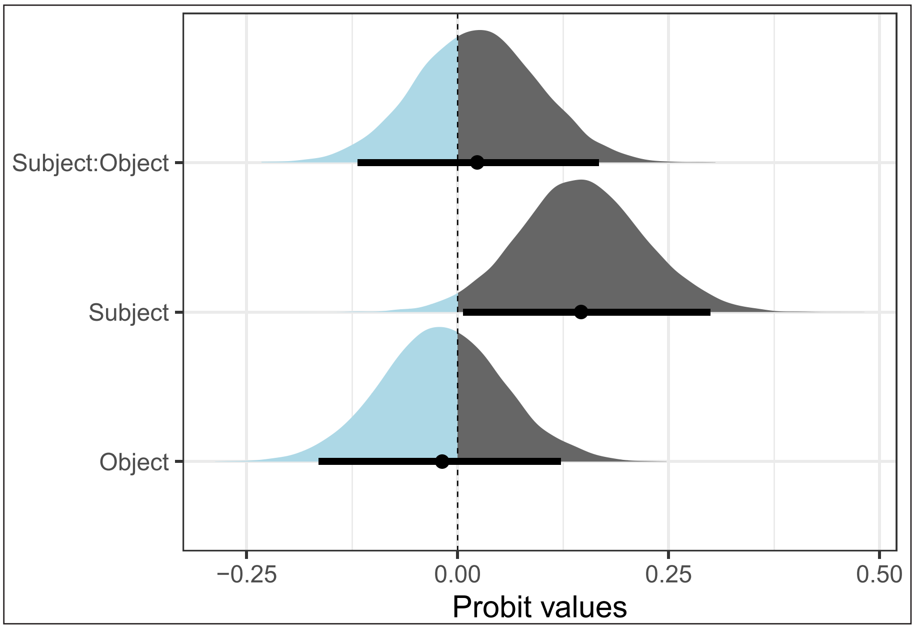


Figure 10: Effects in the acceptability task, Experiment 3b. The dot denotes the mean, the thick lines, the 95% credible intervals. Density areas higher than 0 appear in grey, density areas smaller than 0 appear in blue.

4.3 Experiment 4: Follow-up eye-tracking experiment with only non-referential condition

4.3.1 Participants

We tested 57 participants, of whom 42 self-identified as female. They were recruited via the ILS database of Utrecht University; most of them were students. The mean age was 22.3 (SD: 3.75, range: 18–32). All participants were native speakers of Dutch. None of them reported to suffer from dyslexia, severe eye abnormalities, or other reading problems. Participants with glasses or contact lenses were allowed to participate if their vision was corrected-to-normal. Participants were rewarded with 10 euros.

4.3.2 Materials, design and procedure

The experiment contained 40 target items in only the non-referential conditions, and 100 fillers, of which 40 were actually target items of a different, unrelated experiment with items of comparable length. The number of fillers we used was higher than in the original version, because the target items stood out relatively more because only the non-referential condition was used, which means there were relatively more ‘odd’ items compared to the original experiment. This was balanced out by using more fillers.

An example of one item in all four conditions is shown in (17), where gender of the relevant expressions is marked using subscripts for the original materials in Dutch, and the English translation follows the original. Regions of interest were defined in the same way as in Experiment 2.

(17) **Conditions:**

a. De professor$_{M}$ heeft geen zoon$_{M}$. De laatste jaren werkte hij$_{M}$...
The professor has no son. The last few years he worked... **s.match o.match**

b. De professor$_{M}$ heeft geen dochter$_{F}$. De laatste jaren werkte hij$_{M}$...
The professor has no daughter. The last few years he worked... **s.match o.mis**

c. De professor$_{M}$ heeft geen dochter$_{F}$. De laatste jaren werkte zij$_{F}$...
The professor has no daughter. The last few years she worked... **s.mis o.match**

d. De professor$_{M}$ heeft geen zoon$_{M}$. De laatste jaren werkte zij$_{F}$...
The professor has no son. The last few years she worked... **s.mis o.mis**

...helaas op alle feestdagen.
...unfortunately during all holidays.

Target items were divided over 4 lists via a Latin Square and together with the fillers presented in a unique random order for each participant. Maximally 2 items of the same condition could follow each other. Half of the fillers and half of the target items were followed by a comprehension question, which could be answered by yes or no. Each list started with a practice block, which was the same as in the original version of the experiment.

The procedure was identical to the procedure in the original experiment, as described above. After the eye-tracking experiment, half of the participants took part in the resolution task and half of them in the acceptability judgement task. Because these were both offline tasks, we expected there to be little or no effect of the participants having read the stimuli before in the eye-tracking experiment.

One participant was excluded from the analysis due to poor performance (79% of response questions answered correctly, while all other participants scored above 85%). No participants were excluded based on poor calibration. Two items were excluded from the analysis because the resolution task showed that accommodation was possible for some participants.

4.3.3 Results

As with Experiment 2, we again focus on the results of the critical and the post-critical region in three measures: right-bounded reading times (RB), total fixation durations (TFD), and regression probabilities (RP). The other measures and the wrap-up region do not change the picture presented here. The detailed results of all the measurements are available online.[7]

Critical region: verb + pronoun This is the first region where a match/mismatch of the subject and object with the pronoun can be established. A descriptive summary is shown in **Table 12**. The details of the effects in the critical region are in **Table 13**.

Table 12: Critical region: Mean raw reading times/count summaries by measure and condition.

Subject	**Object**	**RB (ms)**		**TFD (ms)**		**RP (pct)**
		mean	**SE**	**mean**	**SE**	
match	match	289.78	8.05	431.82	13.45	11.58
match	mis	285.42	7.26	439.00	12.88	16.06
mis	match	307.57	8.24	496.10	14.05	12.14
mis	mis	311.32	8.72	525.17	15.73	16.16

[7] https://osf.io/xznw7/.

Table 13: Critical region: Summary of the results. Effects with credible intervals that do not cross 0 are boldfaced.

Measure	Variable	Estimate	Q2.5	Q97.5
RB (log-ms)	**SUBJECT**	**–0.028**	**–0.051**	**–0.005**
	OBJECT	–0.002	–0.027	0.023
	SUBJECT × OBJECT	<–0.001	–0.022	0.022
TFD (log-ms)	**SUBJECT**	**–0.071**	**–0.102**	**–0.039**
	OBJECT	–0.02	–0.054	0.014
	SUBJECT × OBJECT	0.005	–0.022	0.031
RP (log-odds)	SUBJECT	–0.014	–0.196	0.17
	OBJECT	**–0.257**	**–0.488**	**–0.051**
	SUBJECT × OBJECT	–0.017	–0.196	0.161

On RB, a negative effect of SUBJECT was found, indicating a processing cost when the subject gender mismatched the pronoun. A similar effect was found on TFD. This means that when the subject mismatched, participants spent more time in the critical region before leaving the region progressively and spent more time in the region in total. We almost exclusively found effects of SUBJECT in this region on all measures, the one exception being an effect of OBJECT on RP. This effect indicates that regressions from the critical region were more likely when the object's gender mismatched the pronoun.

Post-critical region A descriptive summary is given in **Table 14**. The details of the effects in the post-critical region are summarized in **Table 15**.

Table 14: Post-critical region: Mean raw reading times/count summaries by measure and condition.

SUBJECT	OBJECT	RB (ms)		TFD (ms)		RP (pct)
		mean	SE	mean	SE	
match	match	572.30	18.49	831.84	23.98	28.32
match	mis	549.50	15.96	798.97	22.77	28.82
mis	match	584.72	17.70	893.61	26.05	31.61
mis	mis	613.80	17.12	893.76	23.99	32.95

Table 15: Post-critical region: Summary of the results. Effects with credible intervals that do not cross 0 are boldfaced.

Measure	Variable	Estimate	Q2.5	Q97.5
RB (log-ms)	**Subject**	**–0.037**	**–0.062**	**–0.013**
	object	–0.007	–0.033	0.018
	Subject × Object	0.023	–0.001	0.047
TFD (log-ms)	**Subject**	**–0.054**	**–0.08**	**–0.027**
	Object	0.005	–0.019	0.03
	Subject × Object	0.012	–0.013	0.038
RP (log-odds)	Subject	–0.11	–0.25	0.025
	Object	–0.027	–0.148	0.09
	Subject × Object	0.032	–0.1	0.17

The results in this region are very consistent, in that we almost exclusively found negative effects of subject. The absence of any effect of object supports the hypothesis that non-referential elements are not used during pronoun resolution.

Interestingly, the post-critical region shows a predominantly positive interaction of subject × object on RB. While this interaction crosses 0, it does indicate that a mismatching subject causes a slowdown, and in combination with a mismatching object, this slowdown is even larger. We will come back to this effect in the next section, where we provide a meta-analysis of pooled data from Experiment 2 and Experiment 4.

4.3.4 Discussion

This follow-up experiment with only non-referential quantifiers was carried out in order to increase the power, which may not have been high enough in Experiment 2 to detect an effect of object mismatch. We wanted to find out whether our hypothesis that a mismatching element only causes a slowdown when it is accessible would still be supported by the results. If so, we would expect that only a mismatching subject, which was always accessible in this experiment, would give rise to processing difficulties in the form of longer reading times. An effect of the object, which was always accompanied by a non-referential quantifier in this experiment and was thus inaccessible, would be absent.

In both the critical and the post-critical region and several measures, a clear slowdown was observed when the stereotypical gender of the subject in the preceding sentence mismatched the gender of the pronoun. The results are as would be expected based on previous studies and Experiment 2 in this paper.

There were two effects in which the object played a role. We start with the critical region, where we observed that mismatching objects led to an increase in regressions. This could signal the effect of inaccessible objects on pronoun resolution. However, there are reasons to be skeptical about this interpretation. First, interference effects on retrieval are predicted by models such as Lewis & Vasishth (2005) on reading times, and the effects are commonly observed on reading time measures. Only a few reading studies show retrieval interference effects on regressions (e.g., Appendix A in Jäger et al., 2017, reports only 4 out of 22 studies that show interference effects on regressions; see also Jäger et al., 2020, for discussion). Furthermore, the found effect of OBJECT does not interact with subject match/mismatch, which is surprising from the perspective of cue-based retrieval, and there is no effect of mismatching subject, which is surprising if regressions at this region signaled problems with pronoun resolution. Finally, we note that Experiment 2 did not show a comparable effect. This could be due to the lower power of that experiment, but given that regressions in the same region of that experiment showed a different effect, namely, SUBJECT × OBJECT interaction, it is possible that both effects are accidental. In any case, these findings should definitely be interpreted with caution and probably not linked to pronoun resolution.

In the post-critical region, we saw a predominantly positive interaction of subject and object on right-bounded reading times. This effect could suggest a role for an inaccessible, non-referential, quantifier on pronoun resolution. The effect is observed on early reading measures, in line with several previous studies of interference in retrieval (e.g., Appendix A in Jäger et al., 2017, reports 9 out of 22 studies that show interference effects on early reading measures, first-pass reading times[8]), and it goes in the direction predicted by cue-based retrieval (inaccessible objects cause a slowdown when subjects match, due to cue overload, and inaccessible objects cause a speed-up when subjects mismatch, due to partial match). Right-bounded reading times also show the negative effect of SUBJECT in the post-critical region, which supports the fact that the measure in this region is sensitive to difficulties with pronoun resolution. However, note that the evidence in support of inaccessible objects affecting pronoun resolution remains rather

[8] We mention first-pass reading times as a representative of early reading measures, because right-bounded reading times, which we used, are generally reported and discussed less often in the psycholinguistic literature. Our own preference to present right-bounded reading times in the text stems from the fact that they include information about the direction of reading progress. However, we note that posterior distributions of first-pass reading times closely match right-bounded reading times at this and other regions, and, in particular, they also show a predominantly positive, albeit slightly smaller, interaction. See https://osf.io/xznw7/ for more details on all eye-tracking measures.

uncertain, since the posterior distribution of the relevant parameter crosses 0. Furthermore, it is not clear to what extent this finding is compatible with Experiment 2. For this reason, we merge the data from both eye-tracking experiments and quantify the exact amount of evidence for cue-based retrieval on RB in the following section.

5. Combined analysis of eye-tracking experiments

Since our two eye-tracking studies were similar, we pooled the data from both experiments to investigate how our findings hold in the combined data set. The investigated data set consists of the subset of the first eye-tracking experiment (only data with non-referential quantifiers are used) and the complete second experiment. We focus on two measures: right-bounded reading times and total fixation durations. These measures are studied in two regions: the critical region and the post-critical region. Posterior distributions are summarized in **Table 16**.

Table 16: Analysis of data pooled from two eye-tracking experiments: Summary of the effects. Effects with credible intervals that do not cross 0 are boldfaced.

Region	Measure	Variable	Estimate	Q2.5	Q97.5
Critical	RB (log-ms)	**SUBJECT**	**–0.033**	**–0.052**	**–0.014**
		OBJECT	–0.009	–0.029	0.011
		SUBJECT × OBJECT	–0.003	–0.023	0.016
	TFD (log-ms)	**SUBJECT**	**–0.084**	**–0.110**	**–0.059**
		OBJECT	–0.019	–0.045	0.008
		SUBJECT × OBJECT	–0.002	–0.024	0.020
Post-critical	RB (log-ms)	**SUBJECT**	**–0.039**	**–0.06**	**–0.019**
		OBJECT	–0.008	–0.029	0.013
		SUBJECT × OBJECT	0.018	–0.001	0.039
	TFD (log-ms)	**SUBJECT**	**–0.059**	**–0.08**	**–0.037**
		OBJECT	0.013	–0.007	0.033
		SUBJECT × OBJECT	0.008	–0.013	0.028

In the pooled data, we see a strong negative effect of SUBJECT, showing that across both experiments, mismatching subjects cause a slowdown. There is no clearly negative or positive effect of OBJECT, possibly with the exception of TFD in the critical region, which is predominantly negative but still crosses 0.

Most importantly, we also see a positive interaction of SUBJECT × OBJECT on RB in the post-critical region, which also crosses 0 but is predominantly positive. To further investigate this interaction, we subset data by subject match/mismatch. The results are summarized in **Table 17**. We see that in case of subject match, OBJECT is mostly positive, signalling that matching objects slow down reading. When the subject mismatches, OBJECT is mostly negative, signalling that matching objects cause a speed-up.

Table 17: Analysis of RB data pooled from two eye-tracking experiments, split by subject: Summary of the effects

SUBJECT	Variable	Estimate	Q2.5	Q97.5
Match	OBJECT	0.01	–0.019	0.038
Mismatch	OBJECT	–0.025	–0.056	0.007

Finally, we quantify the amount of evidence for the role of object using Bayes factors. For each estimation of Bayes factor, we consider one measure per region and we subset data by subject match/mismatch and compare two models: the model with OBJECT (the alternative model) compared to the intercept-only model (the null model). The random structure for the two models is identical.

The Bayes factor is sensitive to assumptions about prior distributions (Gelman et al., 2003; Schad et al., 2022). Here, we consider three priors for the effect of OBJECT: the prior distribution centered at zero, with sd values of 0.1 and 0.03; and the prior distribution centered at –0.027, with an sd value of 0.009. The last prior distribution is inspired by Schad et al. (2022), who collect the estimates from the meta-analysis of interference in subject-verb agreement attraction studies in Jäger et al. (2017).

The Bayes factors were estimated using bridge sampling. Since the sampling is variable, we ran each Bayes factor estimation five times and report mean BF_{10} and standard deviation. The values are reported in **Table 18** for subject-mismatch subsets and in **Table 19** for subject match. The calculated Bayes factor BF_{10} shows evidence in favor of the alternative model when its value is greater than 1, and it shows evidence in favor of the null (intercept-only) model when its value is smaller than 1.

Table 18: Bayes factors BF_{10} for pooled data split by subject mismatch, comparing the model with OBJECT and the intercept-only model.

Region	**Measure**	**Prior**	**BF_{10}(mean)**	**BF_{10}(sd)**
Critical	RB	Normal(0, 0.1)	0.15	0.003
		Normal(0, 0.03)	0.45	0.017
		Normal(–0.027, 0.009)	0.43	0.02
	TFD	Normal(0, 0.1)	0.28	0.007
		Normal(0, 0.03)	0.76	0.02
		Normal(–0.027, 0.009)	1.38	0.025
Post-critical	RB	Normal(0, 0.1)	0.55	0.019
		Normal(0, 0.03)	1.27	0.044
		Normal(–0.027, 0.009)	2.98	0.06
	TFD	Normal(0, 0.1)	0.15	0.011
		Normal(0, 0.03)	0.45	0.019
		Normal(–0.027, 0.009)	0.15	0.005

We see that in almost all cases, BF_{10} supports the null model: the model without OBJECT. This strongly reinforces our observation that the inaccessible quantifier is simply ignored for pronoun resolution. There are only two cases in the subject-mismatch subset in **Table 18** that show evidence in favor of the alternative model: total fixation duration in the critical region and right-bounded reading times in the post-critical region. These two cases differ when we consider the subject-match subset; see **Table 19**. The former also shows evidence in favor of the alternative model when we consider the negative prior distribution. In words, the evidence for slowdown due to object mismatch is present irrespective of subject match/mismatch. Right-bounded reading times on the post-critical region, on the other hand, show evidence in favor of the null model for subject-match subsets and for the negative prior distribution. This is compatible with the predictions of cue-based retrieval. However, we should note that even for pooled data and even for very restricted priors, the evidence is very small. The highest value of 2.98 is in the range of anecdotal change in evidence towards the alternative model, according to Jeffreys' interpretation (Jeffreys, 1939).

Table 19: Bayes factors BF_{10} for pooled data split by subject match, comparing the model with OBJECT and the intercept-only model.

Region	Measure	Prior	BF_{10}(mean)	BF_{10}(sd)
Critical	RB	Normal(0, 0.1)	0.18	0.007
		Normal(0, 0.03)	0.52	0.035
		Normal(–0.027, 0.009)	0.63	0.027
	TFD	Normal(0, 0.1)	0.41	0.01
		Normal(0, 0.03)	0.99	0.084
		Normal(–0.027, 0.009)	2.23	0.13
Post-critical	RB	Normal(0, 0.1)	0.18	0.013
		Normal(0, 0.03)	0.52	0.015
		Normal(–0.027, 0.009)	0.1	0.004
	TFD	Normal(0, 0.1)	0.38	0.014
		Normal(0, 0.03)	0.94	0.05
		Normal(–0.027, 0.009)	0.05	0.002

6. General discussion

6.1 Summary of the results

In this study, we investigated the role of semantic accessibility in online discourse processing, and whether or not this can be approached the same way as syntactic accessibility. We did this by presenting a discourse containing a pronoun and multiple possible antecedents, which could match or mismatch the pronoun's gender. Accessibility of these antecedents was manipulated by using a non-referential quantifier, where the target was accessible and the distractor was inaccessible.

The pronoun resolution tasks showed that the distractor was almost never chosen as the resolution for the pronoun, which supports our hypothesis that being inaccessible in discourse blocks an antecedent from being used for pronoun resolution. The acceptability studies showed that, as we expected, conditions without a straightforward resolution for the pronoun (due to inaccessibility and/or gender mismatch) received lower ratings. Conditions with a gender mismatch in the distractor only received lower ratings if the distractor was accessible, which again supports our claim that inaccessible elements are ignored during pronoun resolution.

In the eye-tracking experiments we studied the influence of accessibility and gender mismatch on online processing during pronoun resolution. The results were largely in line with the offline experiments and showed that only the target was used for pronoun resolution; inaccessible distractors did not influence the reading process. Additionally, we pooled comparable data from the eye-tracking experiments and explored Bayes factors comparing models that include or exclude the information about the gender of semantically inaccessible arguments for pronoun resolution. This revealed evidence that was largely in favor of the model that excludes this information, which supports the claim that inaccessible elements are not used during pronoun resolution. There is one exception: early reading measures in the post-critical region showed the strongest evidence for the model that includes the relevant information. This evidence was weak, but it is possible that more data collection will further strengthen it.

6.2 Predictions of cue-based retrieval

In establishing anaphoric dependencies, a suitable antecedent has to be found in the preceding context. The cue-based retrieval model can be used to understand how comprehenders do this in online processing by assuming that establishing such dependencies involves retrieval of memory representations corresponding to potential antecedents. Under this view, comprehenders will have to launch a search in memory when encountering an anaphor to find its antecedent, and this search is typically argued to involve a direct-access cue-matching procedure (Lewis & Vasishth, 2005; McElree et al., 2003; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006). This means that the search directly accesses those items in memory that match the anaphor in retrieval cues, which can consist of features like number or gender.

This model predicts that in a situation in which a grammatically correct target and a grammatically incorrect distractor overlap in cues, this will lead to a cue overload. This cue overload leads to inhibitory interference caused by the distractor, which is associated with a slowdown in reading times. When the target matches all retrieval cues and the distractor does not match any of the retrieval cues, there is no cue overload and hence no inhibitory interference effect. In a situation where a dependency cannot be resolved, for instance, because the target mismatches the necessary cues, a (partially) matching distractor will lead to facilitating interference. This means that processing will be faster when there is a partial match with the cues, compared to when all the cues are mismatching.

In the cue-based retrieval model, all possible antecedents are considered as potential resolutions on the basis of the relevant cues, grammatical or not. Several studies showed that syntactically inaccessible distractors can indeed still cause interference during processing (see, e.g., Badecker & Straub, 2002; Dillon et al., 2013; Jäger et al., 2017, 2020; Kush & Phillips, 2014; Sturt, 2003).

In our study, we approached accessibility from a semantic perspective. Following semantic theories that focus on discourse structure, such as DRT, an element can be accessible or inaccessible in a discourse, and it is only possible to refer to accessible elements. If accessibility in a discourse is similar to syntactic accessibility, the cue-based retrieval model's predictions would remain the same: elements that are inaccessible in a discourse are expected to cause interference in the same way as syntactically inaccessible elements. In our studies, inaccessible elements cause no clearly detectable and convincing interference. Even the strongest evidence for cue-based retrieval, detected in right-bounded reading times in the post-critical region, turned out to be rather weak when quantified using Bayes factors. To some extent, this should not be entirely surprising, since establishing very convincing evidence for cue-based retrieval in syntax is also not entirely straightforward (see Jäger et al., 2017, and Schad et al., 2022, for discussion; see also the next section). However, the lack of clear evidence in both of our reading experiments suggests that discourse might pose even additional challenges to experimental investigations of cue-based retrieval.

6.3 Interference effects in different kinds of dependencies

In the literature testing the predictions made by theories of retrieval, several different types of syntactic constraints were studied. For instance, Kush & Dillon (2021) showed for cataphors that a main subject that mismatched the cataphor in gender caused interference, but only if coreference with the cataphor was syntactically possible, according to Binding Principle B. If this was not the case, the main subject was inaccessible for resolving the cataphor and did not cause interference.

There has been a lot of attention to the comparison of agreement dependencies and reflexive dependencies. Badecker & Straub (2002) found interference due to inaccessible elements in both types of dependencies. Sturt (2003), however, showed for reflexive anaphors that it depends on the stage of processing whether inaccessible elements cause interference or not: in an early stage of processing, he did not find an interference effect, but in a later stage of processing, interference was observed. Dillon et al. (2013) showed that syntactically inaccessible elements in an agreement dependency do cause interference, while they do not lead to interference in a reflexive dependency.

Jäger et al. (2017) performed a meta-analysis of studies on interference in reflexives and subject-verb agreement and non-agreement dependencies, and studied the results in light of cue-based retrieval, as implemented in the account of Lewis & Vasishth (2005) (LV05). Under this account, it is generally assumed that the retrieval mechanism is equal for all syntactic dependencies. In their meta-analysis, Jäger et al. (2017) distinguished full match and partial match situations. They showed that in the case of a full match, there was no clear interference effect for reflexives and subject-verb agreement. Only in non-agreement subject-verb dependencies did they find evidence for the inhibitory effect that would be expected based on the LV05 account. In the case of a partial match, a facilitatory effect as predicted by the LV05 account was only found for subject-verb agreement, while an inhibitory effect was found for reflexives. They argue that the difference found can either be explained as a side-effect of methodological issues, or as a genuine difference in the cognitive processing of the different dependencies, as has already been argued by, a.o., Dillon (2011), Kush (2013), and Xiang et al. (2009). Jäger et al. (2020) performed a Bayesian re-analysis of Dillon et al. (2013) and a large-sample experiment to further investigate the potentially different nature of the interference effect in different types of dependencies. The fact that some studies fail to find an interference effect in reflexives, they argue, can be explained by a lack of statistical power. This claim is supported by the results of their large-sample experiment, where they found interference for both agreement dependencies and reflexive dependencies.

Sturt (2003) used a design comparable to ours, but approached accessibility in terms of syntactic Binding Theory. Only the grammatically correct target could resolve the reflexive anaphor, while in our experiment the target and distractor did not differ in terms of grammaticality; they only differed in terms of accessibility in discourse. In the experiment of Sturt (2003), some participants ended up with an ungrammatical interpretation, while Experiment 1a and 3a showed that the equivalent of this was not the case in our study: the distractor was almost never chosen as a resolution for the pronoun. In addition, in contrast to Sturt (2003), we found no clear interference effect from mismatching inaccessible antecedents in general. However, we observed an effect in early measures (right-bounded reading times), which was supported by the analysis of pooled data, and observed evidence, albeit rather weak, in favor of models with interference. Sturt (2003), in contrast to us, only found an effect in late measures. In sum, the contrast between our findings and those in Sturt (2003) imply that there is a difference in the nature of syntactic and semantic accessibility, although the difference found could also be (partially) attributed to methodological issues and/or a lack of statistical power.

A study that more closely relates to ours is Kush et al. (2015), who studied the effect of inaccessibility on pronoun resolution by comparing referential and quantificational elements. They found that inaccessible elements that *matched* the pronoun's gender did not speed up processing. This is opposite to what could be expected under the LV05 account, where matching properties would facilitate processing in this setting. These findings are in line with our own findings, which suggest that certain kinds of accessibility can prevent potentially interfering elements from affecting processing. Our Experiments 2 and 4 demonstrated that there was no clear effect of a gender mismatch on the inaccessible distractor, while there was such an effect for the target. This suggests that elements that are inaccessible in a discourse do not cause interference associated with cue-based retrieval. In combination with the results of Experiments 1a/b and 3a/b, it can be assumed that during pronoun resolution in discourse, elements that are inaccessible in a discourse are not considered as potential antecedents for the pronoun.

Although the results of the studies discussed above are somewhat mixed, it can be argued that syntactically inaccessible elements can still influence online processing, at least in some stages. This means that the associated syntactic constraints do not play a role in all stages of processing. As a result, syntactically inaccessible elements may be misretrieved, and this leads to interference and processing difficulties. This erroneous retrieval of inaccessible elements is commonly explained by the presence of the cue-based retrieval mechanism, as elements with matching features are retrieved, irrespective of their syntactic properties. This would mean that the cue-based retrieval mechanism can overrule syntactic constraints during (some stages of) processing. However, our findings suggest that elements that are inaccessible in a discourse might not lead to the interference effects expected in the cue-based retrieval model, and that discourse structures thus play an important role in processing. We should add that the exact role of inaccessible elements in discourse hinges on the nature of the (currently very weak) evidence in favor of the model with interference that we observed on right-bounded reading times in the post-critical region in Experiment 2 and in the pooled data. The evidence might signal that there is some genuine, albeit very hard to detect, interference effect due to discourse-inaccessible elements; but it is also possible that the evidence found is just noise. More research into this question is needed to resolve this issue.

In syntax, it is common to define structural constraints which determine whether a sentence is well-formed or not. Although this is less common in semantics, it has been argued that there are constraints that play a role in the construction of discourse. An example of this is the Right Frontier Constraint (see Asher, 1993; Asher & Lascarides, 2003; Polanyi, 1997; Webber, 1988), which states that a discourse constituent must be attached on the right frontier of the ongoing discourse. Discourse accessibility could be seen as a comparable sort of constraint that helps to determine the well-formedness of a discourse. In future work, it should be refined how, and to what extent, the cue-based retrieval mechanism makes use of these types of discourse constraints to establish the correct dependencies in a discourse.

6.4 Other dependencies in a discourse

Our findings in this study on accessibility are based on the resolution of pronouns in a discourse. However, the results may be extended to other types of semantic dependencies, such as presupposition resolution, which is argued to have an anaphoric component (see, e.g., Geurts, 1999; van der Sandt, 1992). In a context such as (18), the presupposition introduced by the presupposition trigger *too* has to be resolved. This means that the discourse has to provide another accessible antecedent who danced with Elisabeth and is not equal to Bill. In this discourse, the presupposition can be resolved to John. In (19), however, this is not possible. As John is introduced in a sub-discourse that is under negation, it is not accessible in the main discourse. This means that the presupposition cannot be resolved, leading to an infelicitous discourse.

(18) John danced with Elisabeth. Bill danced with Elisabeth too.

(19) #John didn't dance with Elisabeth. Bill danced with Elisabeth too.

Another example of a construction where a link between elements in a discourse has to be established is ellipsis, which is also generally assumed to have an anaphoric component (for a review, see Phillips & Parker, 2014). In (20-a), an example of gapping, the verb form in the second clause is elided. This gap can be filled by the antecedent, which is in this case the verb *danced*. The same principle holds for other types of ellipsis, such as sluicing (20-b) and NP-ellipsis (20-c). In all of these cases, inaccessibility of the antecedent would make the discourse infelicitous.

(20) a. John danced with Elisabeth, and Bill ~~danced~~ with Sara.
b. John will pick up Elisabeth, but she doesn't know when ~~he will pick her up~~.
c. John gives the first round, and Bill (gives) the second ~~round~~.

Overall, accessibility plays an important role in constructing a coherent discourse on many different levels. In addition, as we have shown, different types of accessibility have different consequences for processing. However, it is important to note that the resolution process in general differs per type of constraint or dependency. In syntax, irrespective of the type of dependency, it is usually clear what the different constituents are and there is usually a clear one-to-one mapping between them. In semantic dependencies, however, it is much more complicated to establish what the exact constituents are and what role they play in the dependency. As we have seen in the previous examples, for instance, a reference might also refer to a vague concept, such as an event or another (sub)discourse, or it might even entirely disappear from the discourse. This makes a direct one-to-one mapping between the constituents difficult or even impossible. In the end, this means that the processing of syntactic and semantic dependencies will always be inherently different.

7. Conclusion

This study shows that accessibility in a discourse has an effect on pronoun resolution from early on. In two eye-tracking experiments and across early and late measures and in several regions, we did not find convincing evidence that inaccessible antecedents would affect resolution. The strongest evidence in favor of inaccessible antecedents affecting resolution would be classified as anecdotal, according to Jeffreys' scale of evidence.

The fact that inaccessibility in a discourse seems to block an expression from being considered as a potential antecedent is in line with what would be predicted by semantic frameworks that focus on discourse structure, such as DRT. The results contrast with those of studies on syntactic accessibility that showed that syntactically inaccessible antecedents can still influence online processing.

Data accessibility statement

The data analysis and results can be found here: https://osf.io/xznw7/.

Ethics and consent

The studies in this paper have been performed after approval by the Faculty Ethics Assessment Committee Humanities (FEtC-H) of Utrecht University under reference number 20-289-02.

Acknowledgements

The research in this paper was supported by an NWO Vrije competitie grant under grant number VC.GW17.112.

Competing interests

The authors have no competing interests to declare.

Author contributions

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

ORCiD ID's

Tijn Schmitz: 0000-0002-6115-5810; Jan Winkowski: 0000-0002-0301-8471; Morwenna Hoeks: 0000-0002-6734-7370; Rick Nouwen: 0000-0001-9571-4644; Jakub Dotlačil: 0000-0002-5337-8432.

References

Asher, N. (1993). *Reference to abstract objects in discourse* (Vol. 50). Springer Science & Business Media. DOI: https://doi.org/10.1007/978-94-011-1715-9

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

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28* (4), 748. DOI: https://doi.org/10.1037//0278-7393.28.4.748

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

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

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

Carminati, M. N., Frazier, L., & Rayner, K. (2002). Bound variables and c-command. *Journal of Semantics, 19*(1), 1–34. DOI: https://doi.org/10.1093/jos/19.1.1

Chomsky, N. (1981). *Lectures on Government and Binding.* Foris.

Cunnings, I., Patterson, C., & Felser, C. (2014). Variable binding and coreference in sentence comprehension: Evidence from eye movements. *Journal of Memory and Language, 71*(1), 39–56. DOI: https://doi.org/10.1016/j.jml.2013.10.001

Cunnings, I., Patterson, C., & Felser, C. (2015). Structural constraints on pronoun binding and coreference: Evidence from eye movements during reading. *Frontiers in Psychology, 6*, 840. DOI: https://doi.org/10.3389/fpsyg.2015.00840

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reexives. *Journal of Memory and Language, 75*, 117–139. DOI: https://doi.org/10.1016/j.jml.2014.05.006

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

Dillon, B. (2011). *Structured access in sentence comprehension* [Doctoral dissertation, University of Maryland]. DRUM.

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

Dotlačil, J. (2021). Parsing as a cue-based retrieval model. *Cognitive Science, 45*(8), e13020. DOI: https://doi.org/10.1111/cogs.13020

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

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

Geurts, B. (1999). Presuppositions and pronouns. In *Current Research in the Semantics/Pragmatics Interface* (Vol. 3). Brill.

Groenendijk, J., & Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and Philosophy, 14,* 39–100. DOI: https://doi.org/10.1007/BF00628304

Heim, I. (1982). *The semantics of definite and indefinite noun phrases* [Doctoral dissertation, University of Massachusetts, Amherst]. Published 1988: Garland.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language, 94*, 316–339. DOI: https://doi.org/10.1016/j.jml.2017.01.004

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

Jeffreys, H. (1939). *Theory of probability*. Oxford University Press.

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

Kamp, H., & Reyle, U. (1993). *From discourse to logic. Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*. Kluwer. DOI: https://doi.org/10.1007/978-94-017-1616-1

Koornneef, A. W., Avrutin, S., Wijnen, F., & Reuland, E. (2011). Tracking the preference for bound-variable dependencies in ambiguous ellipses and only-structures. In J. Runner (Ed.), *Experiments at the interfaces* (pp. 67–100). Syntax and Semantics series (Vol. 37). Brill. DOI: https://doi.org/10.1163/9781780523750_004

Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing.* (Publication No. 3599956) [Doctoral dissertation, University of Maryland]. ProQuest Dissertations Publishing.

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

Kush, D., Lidz, J., & Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language, 82*, 18–40. DOI: https://doi.org/10.1016/j.jml.2015.02.003

Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology, 5*, 1252. DOI: https://doi.org/10.3389/fpsyg.2014.01252

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

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

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

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

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

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

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

Moulton, K., & Han, C.-h. (2018). C-command vs. scope: An experimental assessment of bound-variable pronouns. *Language, 94*(1), 191–219. DOI: https://doi.org/10.1353/lan.2018.0005

Nicenboim, B., Schad, D., & Vasishth, S. (2021). An introduction to Bayesian data analysis for cognitive science.

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

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science, 42*(S4), 1075–1100. DOI: https://doi.org/10.1111/cogs.12589

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

Parker, D. (2022). Ellipsis interference revisited: New evidence for feature markedness effects in retrieval. *Journal of Memory and Language, 124,* 104314. DOI: https://doi.org/10.1016/j.jml.2022.104314

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language, 94*, 272–290. DOI: https://doi.org/10.1016/j.jml.2017.01.002

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language, 41*(3), 427–456. DOI: https://doi.org/10.1006/jmla.1999.2653

Phillips, C., & Parker, D. (2014). The psycholinguistics of ellipsis. *Lingua, 151*, 78–95. DOI: https://doi.org/10.1016/j.lingua.2013.10.003

Polanyi, L. (1997). Discourse structure and discourse interpretation. *Annual Meeting of the Berkeley Linguistics Society, 23*, 492–503. DOI: https://doi.org/10.3765/bls.v23i1.1263

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2023). *SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading.* arXiv preprint, 2303.05221. DOI: https://doi.org/10.1016/j.jml.2023.104496

Reinhart, T. (1983). *Anaphora and semantic interpretation.* University of Chicago Press.

Reinhart, T., & Reuland, E. (1993). Reflexivity. *Linguistic Inquiry, 24*, 657–720.

Rouder, J. N., Tuerlinckx, F., Speckman, P., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review, 15*(6), 1201–1208. DOI: https://doi.org/10.3758/PBR.15.6.1201

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2022). Workflow techniques for the robust use of Bayes factors. *Psychological Methods.* arXiv preprint, 2103.08744v2.

Stan Development Team. (2020). *RStan: The R interface to Stan.* R package version 2.21.2.

Stan Development Team. (2021). *Stan modeling language users guide and reference manual.*

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language, 48*(3), 542–562. DOI: https://doi.org/10.1016/S0749-596X(02)00536-3

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. DOI: https://doi.org/10.3389/fpsyg.2015.00347

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

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 407. DOI: https://doi.org/10.1037/0278-7393.33.2.407

Van Dyke, J. A., & 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*(3), 285–316. DOI: https://doi.org/10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language, 55*(2), 157–166. DOI: https://doi.org/10.1016/j.jml.2006.03.007

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

Vasishth, S., Bruüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science, 32*, 685–712. DOI: https://doi.org/10.1080/03640210802066865

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

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

Webber, B. L. (1988). Discourse deixis: Reference to discourse segments. In *26th Annual Meeting of the Association for Computational Linguistics* (pp. 113–122). DOI: https://doi.org/10.3115/982023.982037

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language, 108*(1), 40–55. DOI: https://doi.org/10.1016/j.bandl.2008.10.002