Szarvas, Timea. 2025. Why we're still debating principle C reconstruction: A comparative experimental study. *Glossa: a journal of general linguistics* 10(1). pp. 1–42. DOI: https://doi.org/10.16995/glossa.18707

Open Library of Humanities

Why we're still debating principle C reconstruction: A comparative experimental study

Timea Szarvas, University of Potsdam, timea.szarvas@uni-potsdam.de

This paper reports on two experiments on German that investigate the robustness and status of reconstruction of PP modifiers for principle C. Binding, particularly principle C, is a well-established syntactic diagnostic for Ā-movement. However, the facts about principle C reconstruction are unclear – the field has long relied on introspective judgments that contradict one another. Recent experiments on English and German yielded diverging results, but crucially, used very different designs and methods. The contribution of the current paper is an experimental investigation on German principle C reconstruction comparing different experimental designs without changing the lexical material of the items, and as such marks the first methodological study of its kind. One of the two experiments provides evidence for principle C reconstruction: the effect is small and holds for a subset of speakers. The evidence is in line with the broader theory of anaphora resolution where coreference options result from multiple interacting factors, including, but by no means limited to, principle C. I conclude that the principle C reconstruction diagnostic, although valid if employed precisely with a large sample size, is a poor choice to test for the A/Ā-distinction as well as underlying c-command due to the complexity of coreference resolution.

1 Introduction

The existence and robustness of reconstruction for principle C has been subject to a longstanding debate. Nevertheless, it remains a popular diagnostic in the formal syntactician's toolkit – whether the aim is to tease apart A- from Ā-movement (van Urk 2015) or to assess the base position of an element in a complex dependency (Nissenbaum 2000; Citko 2005), principle C reconstruction is prominently featured in the study of syntactic structures and assumed to hold universally across the languages of the world. The empirical picture, however, is rather blurry to say the least. Up until recently, claims were exclusively built on conflicting introspective judgments, and later experimental work has failed to settle the status of the phenomenon. For German, it has been argued that there is reconstruction, but that principle C is a violable constraint (Salzmann & Wierzba & Georgi 2023). Experiments on English have been interpreted both in favor of and against reconstruction (Adger et al. 2017; Bruening & Al Khalaf 2019; Stockwell & Meltzer-Asscher & Sportiche 2021; 2022). The current paper aims to address the discrepancies that have emerged from these recent experiments. For the very first time, the impact of methodological and experimental manipulations on participants' responses to principle C reconstruction are examined, focusing on German.

The paper is structured as follows: Section 2 provides background on principle C reconstruction, including a detailed assessment of the experimental studies so far and how they differ. Section 3 outlines the study on German principle C reconstruction by Salzmann & Wierzba & Georgi (2023) which served as a baseline for the two novel experiments presented thereafter. Section 4 provides a general discussion of the experiments, highlighting the methodological comparison, inter-speaker variability, the implications of the study as well as its limitations. Section 5 concludes.

2 Background

2.1 Principle C reconstruction

This section gives a brief background on principle C reconstruction. Binding principle C states that referring expressions (R-expressions) must not be bound, resulting in disjoint reference between *Poirot* and *he* in (1):

(1) *He_i says that Poirot_i is leaving. (Haegeman 1994: 226)

Although coreference and binding can seem almost identical in many cases, only binding requires that the binder c-command the bindee, while coreference is specified in a discourse model with no syntactic requirements (Reinhart 1983a; b). One environment where binding and coreference yield distinct interpretations is ellipsis (Sag 1976; Reinhart 1983b; Heim & Kratzer 1998).

(2) Gina called her mother. The teacher did, too.
   a. *sloppy reading (binding):* 'The teacher called the teacher's mother.'
   b. *strict reading (coreference):* 'The teacher called Gina's mother.'

In cases with only one potential binder, binding and coreference are indistinguishable. Nevertheless, all else being the same, binding is preferred over coreference (Grodzinsky & Reinhart 1993; Heim 1998). Reinhart (1983b) proposes that only proper binding should be regulated by Binding Theory, a view that has been widely adopted (Büring 2005), with the noteworthy exceptions of Heim (2007) and more recently Bruening (2021). Most of the discussion in the literature, including the claims and data informing the current investigation, relies on a c-command-based definition of binding (though, for this, too, there exist alternative proposals, such as Bruening 2014).

It is maintained that Ā-movement does not create new binding options (Büring 2005). The effect is shown in in (3), where *he* cannot be bound by *a successful athlete* in (3a) but can be bound by *no one* in in (3b). This is taken to indicate that the entire extracted phrase reconstructs to the internal argument position of the verb.

(3) a. *I wonder [whose picture of a successful athlete$_i$] he$_i$ reminded Bill that you saw ___.
b. I wonder [whose comments about him$_i$] no one$_i$ reported ___.
(Sportiche 2017: 16)

On the premise that this is correct, binding can be used to establish A/Ā-distinctions and to probe into the underlying structure of complex dependencies. Notice that in these cases of wh-extraction, it is not the head of the NP itself that causes the violation, but rather a nominal contained in the PP modifier of the wh-phrase. These data are particularly controversial. Some researchers maintain that argumental PP modifiers reconstruct, such as *of a successful athlete* in (3a) and *about him* in (3b). Crucially, PPs functioning as adjuncts do not reconstruct under this view. Those in favor of this distinction cite data such as (4) (van Riemsdijk & Williams 1981; Freidin 1986; Barss 1988; Lebeaux 1988; Sauerland 1998; Takahashi & Hulsey 2009).

(4) a. *Which investigation of Nixon$_i$ did he$_i$ resent ___?
b. Which investigation near Nixon$_i$'s house did he$_i$ resent ___?
(Safir 1999: 589)

The widely adopted explanation for the lack of a reconstruction effect in (4b) is that the adjunct is merged countercyclically after the NP has moved (Freidin 1986; Lebeaux 1988). The adjunct cannot be interpreted in the deeper syntactic position along with the head of the phrase because it has never occupied that position to begin with.[1] Those who disagree with the data in (4), and therefore also (3), argue that nominals cannot take syntactic arguments, meaning that the distinction is merely semantic – syntactically, all PP modifiers are adjuncts, therefore neither of them reconstructs and no principle C violation can be triggered. The reconstructing material,

[1] Note that proponents of the argument-adjunct asymmetry do not unanimously agree with the implementation of Late Merger, criticizing it as being unconstrained and therefore an undesirable solution to the problem (Sportiche 2016; 2019).

under this view, is subject to deletion up to interpretability at LF (Bianchi 1995; Lasnik 1998; Fox 1999; Safir 1999; Kuno 2004; Henderson 2007).

Recall that the validity of the principle C reconstruction test hinges on the assumption that disjoint reference must result from c-command. On the flipside, coreference is taken to be indicative of the absence of a principle C violation. However, we know that several factors can render a principle C violation quite acceptable, such as affectedness (Temme & Verhoeven 2017), interpretive economy, antilogophoricity, processing complexity (Varaschin & Culicover & Winkler 2023), plausibility, salience of the antecedent and at-issueness (Gor 2020). More generally, topichood and contrastive focus (Cowles & Walenski & Kluender 2007; Kaiser 2011), subjecthood (Kaiser 2011), first mention (Järvikivi et al. 2005) and linear proximity (Cunnings & Patterson & Felser 2014) make referents more likely antecedents for a pronoun, showing that anaphora resolution is a highly complex issue going well beyond binding theory. Despite all of these confounds and the conflicting introspective reports, principle C reconstruction remains a popular test to establish A vs. Ā-properties of movement types and c-command relations in complex dependencies. The following section reviews the existing experiments on principle C reconstruction and discusses why they failed to clear up the empirical controversies.

2.2 Previous experimental investigations

We now turn to the experiments conducted so far addressing coreference possibilities under principle C reconstruction. First, I give an overview of the studies and their findings, then we take a look at the specifics in which they differ. Understanding these details is vital to appreciate the effect that they have on the experimental outcomes, which is precisely what the novel experiments in section 3 are designed to address. The existing studies exemplify that multiple roads lead to coreference judgments. Examples (5)–(9) schematically illustrate the different setups.

Adger et al. (2017) employed a forced choice task, collecting data from 91 participants. With positive responses to the alleged principle C violation being at 30%, further increasing with distance, the authors conclude that PP modifiers of nouns do not reconstruct.

(5) Item structure by Adger et al. (2017: 25)

a. Which side of Elizabeth does she prefer ___? (short)

b. Which side of Elizabeth does she say Philip prefers ___? (long)

c. Which side of Elizabeth did Philip say she prefers ___? (longer)

Bruening & Al Khalaf (2019) included two matching referents in the sentence, forcing a choice between them. They compared adjunct vs. argument PP modifiers in surface violations of principle C as well as underlying ones, yielding a 2x2 design. Analyzing data from 75 participants, they report no significant difference between responses for the referent contained in the PP (*the countess*) in the 'wh, adjunct' and 'wh, argument' condition, likewise concluding that PP modifiers do not reconstruct.

(6) Item structure by Bruening & Al Khalaf (2019: 254–255)
   a. The chambermaid told me which portrait of the countess she considered to be the most valuable. (wh, argument)
   b. The chambermaid told me which portrait in the countess's collection she considered to be the most valuable. (wh, adjunct)
   c. The chambermaid told me that she considered one particular portrait of the countess to be the most valuable. (no wh, argument)
   d. The chambermaid told me that she considered one particular portrait in the countess's collection to be the most valuable. (no wh, adjunct)

Stockwell & Meltzer-Asscher & Sportiche (2021), while using items similar to Adger et al. (2017), varied underlying c-command instead of movement by comparing causatives to transitives. They additionally manipulated dependency length, but did not include a comparison between adjuncts and arguments. Participants were asked what the sentence was about, judging the naturalness of two readings on a sliding Likert scale from 0 to 7. One reading indicated coreference (*A picture that Harry framed*) and the other one disjoint reference (*A picture that someone else framed*). Based on the data from 223 participants, the authors conclude that there is reconstruction due to a mean rating of 5.67 for the unnamed referent and 1.95 for *Harry* in the condition 'violation, short'. The significance of the effect vanished in 'violation, long' as well as in 'no violation, short'.

(7) Item structure by Stockwell & Meltzer-Asscher & Sportiche (2021: 206)
   a. Which picture of Harry did he frame ___? (violation, short)
   b. Which picture of Harry did Meghan say he framed ___? (violation, long)
   c. Which picture of Harry ___ made him laugh? (no violation, short)
   d. Which picture of Harry ___ made Meghan say he has good taste? (no violation, long)

In a follow-up study closely matching the previous one, Stockwell & Meltzer-Asscher & Sportiche (2022) included argument vs. adjunct PPs instead of varying the distance of movement. Based on data from 275 participants, they report a mean rating of 2.19 for the coreferent reading for arguments, and a rating of 3.24 for adjuncts in the presence of a violation. This contrast is statistically significant, taken by the authors to indicate that there is an argument-adjunct asymmetry.

(8) Item structure by Stockwell & Meltzer-Asscher & Sportiche (2022: 147)
   a. Which picture of Harry did he frame ___? (violation, argument)
   b. Which picture arranged by Harry did he frame ___? (violation, adjunct)
   c. Which picture of Harry ___ made him laugh? (no violation, argument)
   d. Which picture arranged by Harry ___ made him laugh? (no violation, adjunct)

Salzmann & Wierzba & Georgi (2023) investigated the phenomenon in German with yet another kind of experimental task. As in the experiment by Bruening & Al Khalaf (2019), there were two matching referents in the sentence, but also two yes-no questions per trial. One of them asked whether the sentence could be understood such that the referent in the PP modifier (*embedded referent*, i.e. *Hanna*), in the case of the item presented here, overheard a comment, while the other task asked the same about the referent in the matrix clause (*matrix referent*, i.e. *Lisa*). The relevant measure is the proportion of *yes*-responses to the question about the embedded referent. The authors contrasted the grammatical function of the NP, whether it was moved or in situ, and whether the PP modifier was an argument or adjunct.[2] Based on data from 32 participants, the authors report that coreference with the embedded referent elicited positive reponses in 35.9% of the observations for both the condition 'object, moved, argument' and 'object, moved, adjunct'. The proportion of *yes*-responses to the condition 'subject, moved, argument' and 'subject, moved, adjunct' were at 50.8% and 51.6%, respectively. The authors report a significant effect of PHRASE, but no significant effect of the argument-adjunct distinction, concluding that both arguments and adjuncts reconstruct in German. Relatively high coreference rates led to the conclusion that principle C is violable in German.

(9) Item structure by Salzmann & Wierzba & Georgi (2023: 602–603)

a. Object, moved, argument

Lisa	erzählt,	welche	Geschichte	über	Hanna	sie	___
Lisa-NOM	recount-3SG	which	story-ACC	about	Hanna-ACC	she-NOM	
ärgerlich	fand.						
upsetting	find.PST.3SG						

'Lisa recounts which story about Hanna she found upsetting.'

b. Object, moved, adjunct

Lisa	erzählt	welche	Geschichte	im	Buch	über	Hanna
Lisa-NOM	recount-3SG	which	story-ACC	in.DAT	book-DAT	about	Hanna-ACC
sie	___ ärgerlich	fand.					
she-NOM	upsetting	find.PST.3SG					

'Lisa recounts which story in the book about Hanna she found upsetting.'

c. Subject, moved, argument

Lisa	erzählt,	welche	Geschichte	über	Hanna	___	sie
Lisa-NOM	recount-3SG	which	story-NOM	about	Hanna-ACC		she-ACC
verärgert	hat.						
upset	have.PST.3SG						

'Lisa recounts which story about Hanna has upset her.'

[2] Due to space restrictions, the in situ conditions are omitted from (9). They featured embedded *that*-clauses.

d. Subject, moved, adjunct
Lisa erzählt, welche Geschichte im Buch über Hanna ___
Lisa-NOM recount-3SG which story-NOM in.DAT book-DAT about Hanna-ACC
sie verärgert hat.
she-ACC upset have.PST.3SG
'Lisa recounts which story in the book about Hanna has upset her.'

To summarize, two groups argue that PP modifiers never reconstruct in English, one group argues that argumental PP modifiers do, and the group studying German argues that all types of PP modifiers reconstruct, concluding that it is principle C rather than reconstruction that is compromised. Despite aiming to test the same phenomenon, the experimental data are not straightforwardly comparable due to the use of different methods. Further confounding factors include the experimental design, item complexity and structure. It is also reasonable to assume that researchers interpret the data differently. While syntactic theory makes categorical predictions, the results obtained do not depict these extremes. Experiments show that participants respond positively to a supposed principle C violation in 20–50% of the cases. Speakers therefore seem to have tendencies at most, which is expected based on reports of non-syntactic factors making principle C violations acceptable (Temme & Verhoeven 2017; Varaschin & Culicover & Winkler 2023; Gor 2020), and work arguing for the importance of linear order in binding (Bruening 2014). The subsequent section critically reviews the limits of what we can conclude based on the data so far.

2.3 Evaluation of previous experiments

2.3.1 Methodological differences

We now delve into the differences between the previous studies with respect to the methods they have employed. Adger et al. (2017) collected yes-no responses to a question asking about coreference with the only matching referent in the sentence. They did so by highlighting the R-expression and the pronoun, and asking if participants could use the sentence when the two referred to the same individual. Sentences were presented without context. Only those participants who passed a test where they had to indicate the impossibility of coreference with surface-level principle C violations were included in the sample.

Bruening & Al Khalaf (2019) collected a forced choice response between the two matching referents in the sentence, measuring which referent was preferred among participants. Although we cannot tell whether participants truly disallow a referent, the authors argue that the preference-based task is fairly simple and natural. If a grammatical constraint rules out either referent, responses should be close to zero, while chance level performance would indicate no such constraint. It is acknowledged that other factors, such as word order and the alternation between cataphoric and anaphoric reference may distort the picture further. The authors mention

that rates should be ‘significantly different from zero’ if there is no reconstruction, though it is unclear what exactly counts as significantly different. Sentences were presented without context in this experiment, too.

Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) used a sliding Likert scale for each intended reading ranging from 0 to 7, arguably the most complex method. In the relevant condition, participants had to judge the naturalness of the reading violating principle C and of the reading where the pronoun referred to an unknown referent. There were thus two independent tasks per trial. In this experiment, too, sentences were presented without specific contexts, but participants were informed before the experiment that they should imagine being at a party and joining an ongoing conversation. Although the aim of this complex task was likely to allow for more nuanced responses, the authors report that the mean ratings of the two readings in each condition add up roughly to 8.0. This indicates that an increased rating for one reading led to a decreased rating for the other reading. Rather than improving the precision of the measurement, it appears that the complexity encouraged people to engage in simplification strategies of their own.

In the experiment by Salzmann & Wierzba & Georgi (2023), participants had to complete two forced choice tasks. The two questions were whether the sentence could be interpreted such that (i) the referent in the matrix clause could corefer with the pronoun or (ii) the referent in the PP could corefer with the pronoun, respectively. The intention was to collect explicit judgments on both interpretive options. In addition, the authors informed participants that the sentences may have multiple interpretations, even if one of them is more readily available than the other, and that they should consider each of the options carefully. In a training sequence before the experiment, participants were shown an ambiguous sentence to illustrate this point (*Maria hat Anna besucht, weil sie nett ist* ‘Mary visited Anna because she is nice’). Though other groups do not report about their instructions in as much detail, it appears that Salzmann & Wierzba & Georgi (2023) explicitly communicated the intention of the task to their participants, which on the one hand, ensures that they understood it, but on the other hand, may also encourage participants to spend more time thinking about the items and being less spontaneous with their judgments. In general, there is nothing inherently wrong about instructing participants more explicitly or inviting them to think about their responses. However, in the case of coreferent judgments, which are known to be highly susceptible to pragmatic factors, participants may end up accepting the coreferent reading regardless of a syntactic violation if they take enough time to contemplate.

While there is some level of uncertainty associated with giving people only one forced choice task, providing them with more complex tasks and instructing them to consider each option carefully may induce a bias to be more accepting and potentially consider non-syntactic cues that may facilitate the acceptability of a violation. Given the involvement of non-syntactic factors in coreference resolution, these factors may have additionally influenced the outcomes.

2.3.2 Item structure

Let us now turn to the differences between the experimental items themselves across studies. Authors vary with respect to whether they presented participants with interrogative or declarative sentences embedding the wh-dependency. This coincides with the presence of an alternative referent that the pronoun could saliently refer to.

(10) a. Which side of Elizabeth does she prefer? (Adger et al. 2017)
b. The chambermaid told me which portrait of the countess she considered to be the most valuable. (Bruening & Al Khalaf 2019)
c. Which picture of Harry did he frame? (Stockwell & Meltzer-Asscher & Sportiche 2021; 2022)
d. Lisa erzählt, welche Geschichte über Hanna sie ärgerlich fand.
Lisa-NOM recount-3SG which story-ACC about Hanna-ACC she-NOM upsetting find.PST.3SG
'Lisa recounts which story about Hanna she found upsetting.' (Salzmann & Wierzba & Georgi 2023)

Notice the three-way opposition in previous experiments regarding alternative referents: in the case of Adger et al. (2017), the possibility of an alternative referent was not mentioned at all, Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) gave participants scales to assess coreference with the referent in the sentence and *someone else*, while Bruening & Al Khalaf (2019) and Salzmann & Wierzba & Georgi (2023) included a named alternative referent in their experimental items. If not given an alternative at all or if the alternative is not particularly salient, participants may be biased to resolve pronominal reference with whatever referent is available (Gordon & Hendrick 1998). It is therefore unclear if, in the case of Adger et al. (2017), fairly high coreference rates were obtained because there was no reconstruction, or if people simply chose the referent present in the sentence because it is less costly than postulating another referent in the discourse model without any prompt to do so. On the other hand, the properties of the alternative referent included by Bruening & Al Khalaf (2019) and Salzmann & Wierzba & Georgi (2023) may have the opposite effect. In both experiments, the referent was the subject of the matrix clause, making it a strong contender for coreference and a potential distraction from the embedded referent. Participants may have responded less favorably to the referent in the PP simply because the alternative referent was extremely prominent.

There are also some potential confounds resulting from the choice of verbs. In particular, Salzmann & Wierzba & Georgi (2023) used psych verbs to maintain minimal pairs across conditions with subject and object extraction, allowing for pairs such as *which story about Hanna delighted her* and *which story about Hanna she found delightful*. Temme & Verhoeven (2017) argue

that affectedness can increase the acceptability of principle C violations particularly in German, meaning that higher coreference rates should be obtained with psych verbs than transitive verbs. However, in that case one may find that varying verb type across different conditions of the same item plays a role, and arguably, this would lead to greater variability across items than consistently using verbs with potential confounds. The next section discusses the experimental designs, in particular the control construction(s) that the supposed underlying principle C violation has been compared to in the existing studies.

2.3.3 Experimental designs

We now examine the experimental designs employed by previous researchers in more detail. Adger et al. (2017) did not collect data for PP modifiers functioning as adjuncts, only arguments, given that reports in the literature converge about the former but not the latter. All authors collected responses to surface violations of principle C, intended as a baseline as to what response a violation should elicit.[3] Salzmann & Wierzba & Georgi (2023); Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) additionally used items where only underlying c-command was manipulated without changing the linear order of the R-expression and the pronoun. Salzmann & Wierzba & Georgi (2023) varied the grammatical function of the extracted phrase, while Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) used causative constructions such as *which picture of Harry made him laugh*, comparing them to regular transitives to achieve this distinction.

The divide between the groups arguing for and those arguing against the reconstruction of PP modifiers corresponds to the use of c-command as an experimental factor. Recall that Salzmann & Wierzba & Georgi (2023); Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) argue that there is reconstruction while Adger et al. (2017); Bruening & Al Khalaf (2019) argue that there is not. In the former groups' experiments, items without an underlying principle C violation allow for coreference more readily than ones where reconstruction leads to a violation. The experiments by Adger et al. (2017); Bruening & Al Khalaf (2019) base their arguments on the observation that supposed underlying violations are indicated to be possible much more frequently than surface violations of principle C, which, however, is confounded with the linear order of the pronoun and the R-expression (via controls in the case of Adger et al. 2017). The evidence provided by Stockwell & Meltzer-Asscher & Sportiche (2021; 2022); Salzmann & Wierzba & Georgi (2023), on the other hand, provides straightforward information on the relevance of c-command. The review of experimental designs in this section highlights that previous claims have been made on quite disparate grounds, making the experiments difficult to evaluate side by side. This is why interpreting the findings and placing them into context is particularly important. In what follows, we take a closer look at the different ways in which previous authors chose to do this.

[3] In the case of Adger et al. (2017) and Stockwell & Meltzer-Asscher & Sportiche (2021; 2022), movement was not included as an experimental factor, supplying merely descriptive statistics for these items.

2.3.4 Interpreting the results

To contextualize previous authors' conclusions, it is crucial to understand how they interpret their results. Groups measuring proportions of responses unanimously report them to be below chance for the condition in which a principle C violation is predicted to occur, yet interpret this below-chance performance in different ways. The groups arguing against reconstruction, i.e. Adger et al. (2017) and Bruening & Al Khalaf (2019), rely on the observation that responses indicating the possibility of or the preference for coreference between the pronoun and the R-expression are well above zero. They argue that if there was a grammatical constraint, these kinds of results would be unexpected, especially because they found that responses were below 10% in items where a principle C violation without movement is involved. The authors therefore conclude that PP modifiers do not reconstruct based on a contrast between items with and without movement. Salzmann & Wierzba & Georgi (2023) also find coreference rates well above zero but below chance in conditions with PP modifiers functioning as arguments (35.9%). Crucially, however, they reach a different conclusion due to separating linear precedence from underlying c-command by comparing object and subject extraction (instead of surface vs. underlying violations of principle C, where a null result would indicate full reconstruction). Because coreference rates differ by roughly 15% between the conditions with and without an underlying principle C violation, and this effect is statistically significant, they conclude that there is reconstruction. They further argue against an argument-adjunct asymmetry in German because the exact same coreference rate was obtained with argument and adjunct PPs. We see, therefore, that similar coreference rates are interpreted in different ways depending on what comparisons research groups have included. Designs relying on a null effect, such as Adger et al. (2017) and Bruening & Al Khalaf (2019), are difficult to evaluate given that one has to rely on an arbitrary threshold-based reasoning (for a thorough discussion, see Salzmann & Wierzba & Georgi 2023).

Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) collected judgments on a scale for each available reading rather than a binary coreference judgment. In their first experiment where the effect of distance was tested with PP arguments, they report that the coreferent reading scored a mean rating of 1.95, while the reading where the pronoun referred to an unnamed referent scored a mean rating of 5.67. They argue in favor of reconstruction because the rating for the unnamed referent increases if a violation is present. In their second experiment, the mean rating for the coreferent reading with PP arguments was 2.19 and 3.24 for PP adjuncts. Because this contrast is statistically significant, the authors argue that there is an argument-adjunct asymmetry in English. Recall that this group also isolated c-command from linear order by comparing causative and transitive structures. The experiment revealed a higher mean rating for the coreferent reading in the absence of a principle C violation, leading the authors to conclude that PP modifiers do reconstruct.

Recall that Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) hypothesize that participants may have responded with preferences rather than possibilities. This is supported by the observation that the mean ratings of the two readings in each condition add up to roughly 8.0, indicating that an increase of one reading's rating led to the decrease of the other's rating. Notice also that the authors interpret an increased rating for the alternative referent as an indicator for a principle c violation rather than focusing on whether the rating for the embedded referent decreased. The definition of a principle C effect is the unavailability of coreference with the R-expression that is c-commanded by the pronoun, not the prominence of alternatives. The results of Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) thus may be quite similar to those obtained by Bruening & Al Khalaf (2019), but the groups crucially differ in how they interpret them. For Stockwell & Meltzer-Asscher & Sportiche (2021; 2022), determining whether there is a principle C reconstruction effect is based on two components: the decrease of ratings for the named referent as well as the increase of ratings for the unnamed referent. That is, while mean ratings for the coreferent reading may look similar in the argument and adjunct conditions (reflected in the lack of a significant main effect of RELATION reported by Stockwell & Meltzer-Asscher & Sportiche 2022), it is crucially the larger increase of the mean rating for the unnamed referent that leads the authors to the conclusion that there is an argument-adjunct asymmetry (the significant interaction of RESPONSE and RELATION). For Bruening & Al Khalaf (2019), and arguably also Adger et al. (2017), the possibility of coreference with the R-expression in the PP being well above zero is taken to indicate that there cannot be a syntactic violation. These differences in what researchers decide to focus on reveal a fundamental problem – syntactic theory makes categorical predictions, yet the data do not reflect them. It is up to the researcher to tease apart the confounds and identify the confounding factors that may lead to the observed results. The fact that in most studies, a threshold-based reasoning needs to be adopted to interpret the data creates problems due to the lack of objective parameters.

Running experiments is not only a welcome addition to introspective data, it is necessary if the introspective judgments diverge. However, if supposedly objective, experimental evidence seems to further fuel the debate rather than settle it, we need to investigate whether the experimental design and methods are causing this. The existing experiments, when compared, reveal a lot about what to consider when studying principle C reconstruction experimentally. However, they do not provide a definitive answer to whether principle C reconstruction exists. I came up with two experiments following up on the work by Salzmann & Wierzba & Georgi (2023), focusing on the experimental task and the presence of an alternative referent in the target item, while keeping the lexical content of the items consistent. By doing so, the novel data allow us to verify how much of the discrepancy between the results stems from actual properties of principle C reconstruction, and how much is indebted to methodological and item-related confounds.

3 Experiments

3.1 Baseline

Before presenting the novel experiments on principle C reconstruction in German, I discuss the baseline study by Salzmann & Wierzba & Georgi (2023) in more detail. Although the following investigation is on German, given that the manipulations are rather methodological and not language-specific, the findings should hold cross-linguistically. To reiterate briefly, one of the experiments by Salzmann & Wierzba & Georgi (2023) investigated whether there was an argument-adjunct asymmetry in the reconstruction of PP modifiers, which is the experiment discussed herein. The authors manipulated the factors MOVEMENT (moved vs. in situ), PHRASE (subject vs. object) and STATUS (argument vs. adjunct). The relevant conditions for the current matter are 'moved, subject, argument' and 'moved, object, argument' – focusing on the methods, the argument-adjunct asymmetry will not be included in the experiments. Each target item contains a wh-extracted NP with a PP argument modifying it. The PP contains an R-expression matching the features of the pronoun that linearly follows it in the sentence. This extraction dependency is embedded into a matrix clause containing another matching R-expression. The example in (11) repeats the item structure of the two conditions discussed hereafter:

(11) Item structure by Salzmann & Wierzba & Georgi (2023)

a. Object, moved, argument

Lisa	erzählt,	welche	Geschichte	über	Hanna	sie	___
Lisa-NOM	recount-3SG	which	story-ACC	about	Hanna-ACC	she-NOM	
ärgerlich	fand.						
upsetting	find.PST.3SG						

'Lisa recounts which story about Hanna she found upsetting.'

Can the sentence be understood such that…

Lisa found a story upsetting? ☐ yes ☐ no

Hanna found a story upsetting? ☐ yes ☐ no

b. Subject, moved, argument

Lisa	erzählt,	welche	Geschichte	über	Hanna	___	sie
Lisa-NOM	recount-3SG	which	story-NOM	about	Hanna-ACC		she-ACC
verärgert	hat.						
upset	have.PST.3SG						

'Lisa recounts which story about Hanna has upset her.'

Can the sentence be understood such that…

A story has upset Lisa? ☐ yes ☐ no

A story has upset Hanna? ☐ yes ☐ no

Using a Latin square design, the experiment tested 32 items in eight conditions in addition to 44 distractors. Items were not accompanied by a context sentence. 32 native speakers of German were recruited over Prolific. Participants were instructed to read each item carefully. They were also explicitly instructed to consider both of the interpretations offered, noting that in some cases, both may be possible. Items were presented in a randomized order. Two forced choice questions were shown per sentence, one assessing coreference with the matrix referent and the other with the embedded referent. The order of the two questions was likewise shown in a randomized order. **Table 1** shows the results.

Phrase	Arg/adj	Movement	Q matrix	Q embedded
object	argument	in situ	97.8%	7.0%
object	argument	moved	94.5%	35.9%
object	adjunct	in situ	96.9%	7.8%
object	adjunct	moved	96.9%	35.9%
subject	argument	in situ	93.0%	65.6%
subject	argument	moved	95.3%	50.8%
subject	adjunct	in situ	96.1%	55.5%
subject	adjunct	moved	96.1%	51.6%

Table 1: Results of experiment 2 reported in Salzmann & Wierzba & Georgi (2023: 38), conditions tested in experiments reported herein are highlighted.

The authors report a significant simple effect of PHRASE within the levels moved and argument. Coreference rates with the referent in the PP illustrated by **Figure 1**, meaning the proportion of *yes*-responses to the respective question, were 35.9% in the object, argument, moved (11a) and 50.8% in the subject, argument, moved condition (11b). The proportion of positive responses to the question about coreference with the matrix referent was above 90% across all conditions, indicating that participants understood the task. The authors take the findings to indicate that PPs functioning as arguments reconstruct, acknowledging that surface position matters as well based on an effect of MOVEMENT.

The authors include a discussion of why coreference rates are neither close to 0% nor to 100%, just like in the experiments on English. They speculate that non-syntactic factors may be at play, addressing two prominent accounts in follow-up experiments. One of them assessed whether parallel function of the R-expression and the pronoun had an effect, the other whether subjects were particularly prominent referents. Neither of the experiments yielded a significant outcome. The authors therefore conclude (i) that there is reconstruction for PP modifiers, and (ii) that principle C informs coreference possibilities, although it is a violable constraint in German.

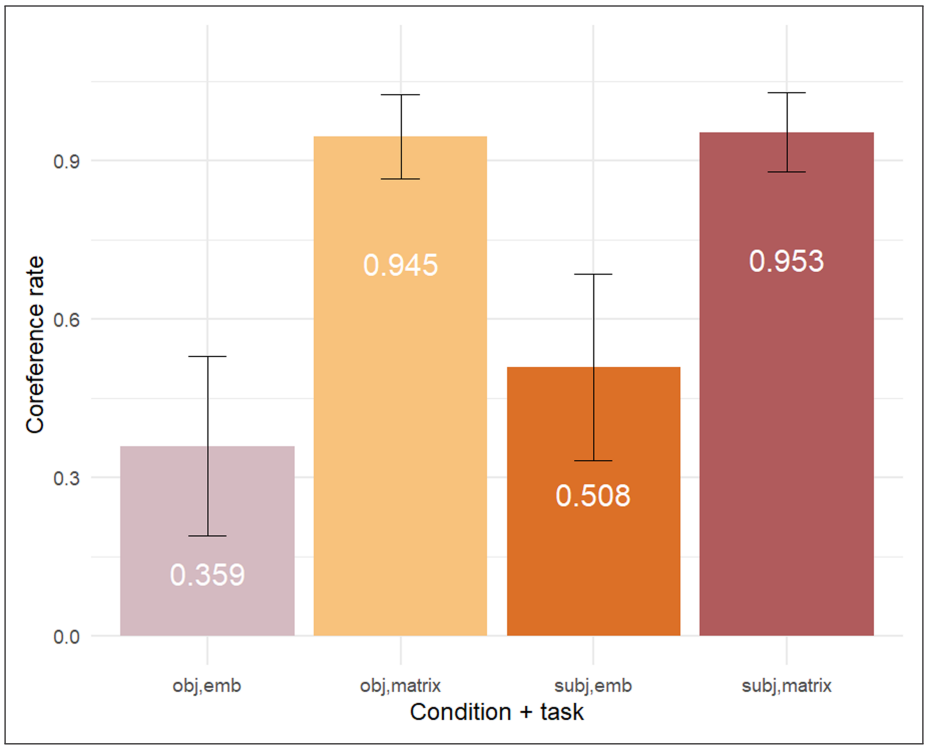

Figure 1: Coreference rates across conditions in experiment 2 reported in Salzmann & Wierzba & Georgi (2023). Error bars indicate standard error.

There are some remaining issues, however, that receive no straightforward explanation. The contrast between the conditions 'object, argument, moved' and 'subject, argument, moved' is arguably low at 14.9%. If the source of the contrast is a principle C violation under reconstruction, one would expect a contrast similar to the one found between 'object, argument, in situ' and 'subject, argument, in situ', which is at 58.6%. The authors note that participants may vary with respect to how much they prefer anaphoric over cataphoric reference, since the pronoun precedes the referent in in situ conditions, but follows it in moved conditions. However, this is in conflict with the conclusion that reconstruction is robust and that it is principle C that is violable in German – the effect of the principle C violation in in situ conditions is much more pronounced than in moved conditions.[4] Further, no difference in coreference rate is obtained between arguments and adjuncts in the reconstructing conditions. This is in line with the view that there is no argument-adjunct asymmetry. However, note that those arguing against the asymmetry maintain that neither adjuncts nor arguments should show reconstruction effects, while the authors conclude the opposite, arguing that both types of PPs reconstruct in German. This entails

[4] Nonetheless, linear order and c-command coincide in the in situ conditions, making it difficult to determine which of the two produces the contrast.

cross-linguistic variability. Note also that the coreference rates in the 'subject, in situ' conditions are still fairly low at 65.6% and 55.5%, respectively, despite the absence of a syntactic violation. They crucially also differ quite substantially despite no difference in linear and structural precedence as well as c-command. Syntactic theory predicts equal availability of coreference in both of these conditions. This demonstrates that cataphoric reference is dispreferred regardless of a violation.

Salzmann & Wierzba & Georgi (2023) designed the instructions, task and item structure with a lot of attention to detail. While the aim was to eliminate ambiguity both for the participants as well as the interpretability of the data, the level of precision made the experiment particularly demanding. Note that participants were presented 78 items with two questions each. The results may be skewed if participants became inattentive in later trials. Furthermore, the high saliency of the matrix referent suggests that it may have distracted participants from the embedded referent, introducing another unwanted bias in the experiment. To investigate these issues in particular, two experiments were designed constituting simplified versions of the study by Salzmann & Wierzba & Georgi (2023). The first experiment, which we turn to in the following section, departs from the original by including only one forced choice task per trial.

3.2 Experiment 1: simplification of the task

3.2.1 Materials

In the first experiment, the 32 items from experiment 2 by Salzmann & Wierzba & Georgi (2023) were used, including only the conditions 'object, argument, moved' and 'subject, argument, moved'.

(12) Context:
Bei Lisas und Hannas Familienfest wurden peinliche Geschichten aus
at Lisa's and Hanna's family.party AUX.PST.PASS embarrassing-PL story-PL from
ihrer Kindheit erzählt.
3PL.POSS.DAT childhood tell.PTCP
'At Lisa's and Hanna's family reunion, embarrassing stories from their childhood were told.'

a. Object, matrix
Lisa erzählt, welche Geschichte über Hanna sie ___
Lisa-NOM recount-3SG which story-ACC about Hanna-ACC she-NOM
ärgerlich fand.
upsetting find.PST.3SG
'Lisa recounts which story about Hanna she found upsetting.'
Lisa found a story upsetting. ☐ yes ☐ no

b. Subject, matrix
Lisa erzählt, welche Geschichte über Hanna ___ sie verärgert hat.
Lisa-NOM recount-3SG which story-NOM about Hanna-ACC she-ACC upset have.PST.3SG
'Lisa recounts which story about Hanna has upset her.'
A story has upset Lisa. ☐ yes ☐ no

c. Object, embedded
Lisa erzählt, welche Geschichte über Hanna sie ___ ärgerlich fand.
Lisa-NOM recount-3SG which story-ACC about Hanna-ACC she-NOM upsetting find.PST.3SG
'Lisa recounts which story about Hanna she found upsetting.'
Hanna found a story upsetting. ☐ yes ☐ no

d. Subject, embedded
Lisa erzählt, welche Geschichte über Hanna ___ sie verärgert hat.
Lisa-NOM recount-3SG which story-NOM about Hanna-ACC she-ACC upset have.PST.3SG
'Lisa recounts which story about Hanna has upset her.'
A story has upset Hanna. ☐ yes ☐ no

Additionally, 24 target items from another experiment exploring a different research question were presented. Eight unambiguous fillers were also included. Before testing started, participants were guided through three training trials, including feedback explaining the logic of the task. Targets items were distributed in a 2x2 Latin square design.

3.2.2 Method

Participants were presented a single coreference judgment task per trial. Participants only had to decide whether the pronoun could refer to either of the referents, varying which referent the task was about across trials. Metalinguistic terms were avoided by repeating the sentence with the intended reading (cf. Salzmann & Wierzba & Georgi 2023). Each sentence was accompanied by a neutral context introducing both referents, aiming to mitigate the prominence of the matrix referent by having both referents present in the discourse before encountering them in the target item. Whether the wh-phrase was an object or a subject was encoded as the factor PHRASE. This factor was fully crossed with another factor, REFERENT, encoding whether the task inquired about the matrix or the embedded referent, with the matrix conditions serving as a sanity check. In

subject conditions, both the surface and the base position of the extracted phrase c-commands the pronoun, i.e. there should be no principle C violation regardless of reconstruction, and coreference between the embedded referent and the pronoun should always be possible. In object conditions, however, the surface position of the displaced phrase c-commands the pronoun, but under reconstruction, this relation is reversed, i.e. the pronoun c-commands the base position of the displaced phrase, yielding a principle C violation.

3.2.3 Participants

A total of 175 participants were recruited over the platform Prolific, with mean age 32.09, n = 87 identifying as male, n = 84 as female, and n = 4 as non-binary. Monetary compensation was provided for everyone who completed the experiment successfully. Participants were native speakers of German living and raised in Germany (76.3%), Austria (19.08%) or Switzerland (4.62%). The experimental conditions 'object, matrix' and 'subject, matrix' served as attention checks, with failure to indicate coreference with the matrix referent in more than 20% of cases leading to exclusion. This threshold was also implemented to assess whether participants understood the task and considered coreference possibilities rather than preferences. Based on this criterion, 25 participants were excluded.

3.2.4 Procedure

The experiment was set up through the platform L-Rex (Starschenko & Wierzba 2024). Participants were instructed to read the items carefully, but to decide based on their first impression. They were also encouraged to use the comment function in case they felt that the forced choice task alone did not fully express their judgment. The sentences were displayed simultaneously with the context, the latter in italics. The question was shown below the sentence with the answer options *yes*, indicating coreference between the respective referent and the pronoun, and *no*, indicating disjoint reference. The first block included the training items in non-randomized order, first showing an example of coreference, then disjoint reference, and an ambiguous sentence as the third example. Participants were explicitly guided through this training sequence and were given feedback indicating that *yes* should be chosen in cases of ambiguity. This was done to ensure that the experiment assessed coreference *possibilities*, not *preferences*. The second block contained target items, pseudofillers and fillers in pseudo-randomized order, such that two items from the same set of materials were never shown consecutively.

3.2.5 Hypotheses

Assuming that the mediocre proportions of coreference reported by Salzmann & Wierzba & Georgi (2023) are due to the complexity of the experimental task, and that their conclusions about reconstruction are correct, the majority of responses in the condition 'object, embedded' should be *no*. Responses indicating coreference with the referent in the sentence are expected to occur in 25% or less of the observations in 'object, embedded'. Simultaneously, the majority of

responses in the condition 'subject, embedded' should be *yes*, given that there is no principle C violation. Responses indicating the possibility of coreference are expected to occur in around 75% of the observations, indicating above chance performance. Neither floor nor ceiling effects were expected in either condition due to non-syntactic influences on coreference.[5] Overall, simplifying the experimental task is expected to yield clearer results than found in previous work. This should manifest in a significant main effect of PHRASE if principle C reconstruction is a robust syntactic constraint based on c-command. On the other hand, if PP modifiers do not reconstruct, we expect no pronounced difference between the conditions 'object, embedded' and 'subject, embedded', meaning no significant effect of PHRASE.

3.2.6 Results

Coreference rates across conditions are illustrated in **Figure 2**. Participants chose the answer *yes*, indicating the possibility of coreference, in 19.4% of the observations in the condition 'object, embedded', and in only 26.4% of the observations in the condition 'subject, embedded'. At the same time, coreference rates in conditions inquiring about the matrix referent are between 96.1% and 98.8%.

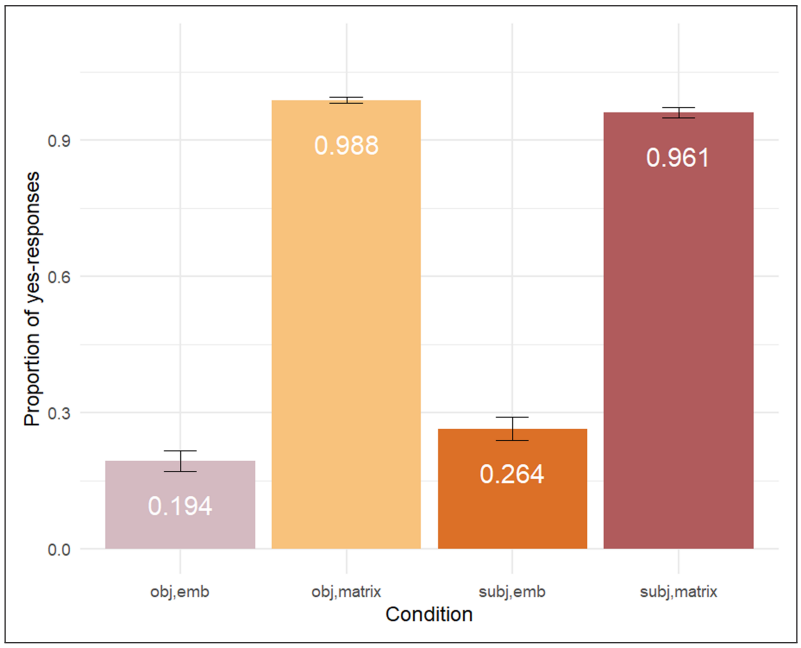

Figure 2: Coreference rates across conditions in experiment 1. Error bars indicate standard error.

[5] The justification for the cut-off points at 25 and 75% are based on the previous study by Salzmann & Wierzba & Georgi (2023), who found coreference rates around 35–50%. Since the aim of the study was to amplify the contrast between the conditions, the expectation was that coreference rates would be pushed further toward the extremes depending on the presence or absence of a principle C violation under reconstruction.

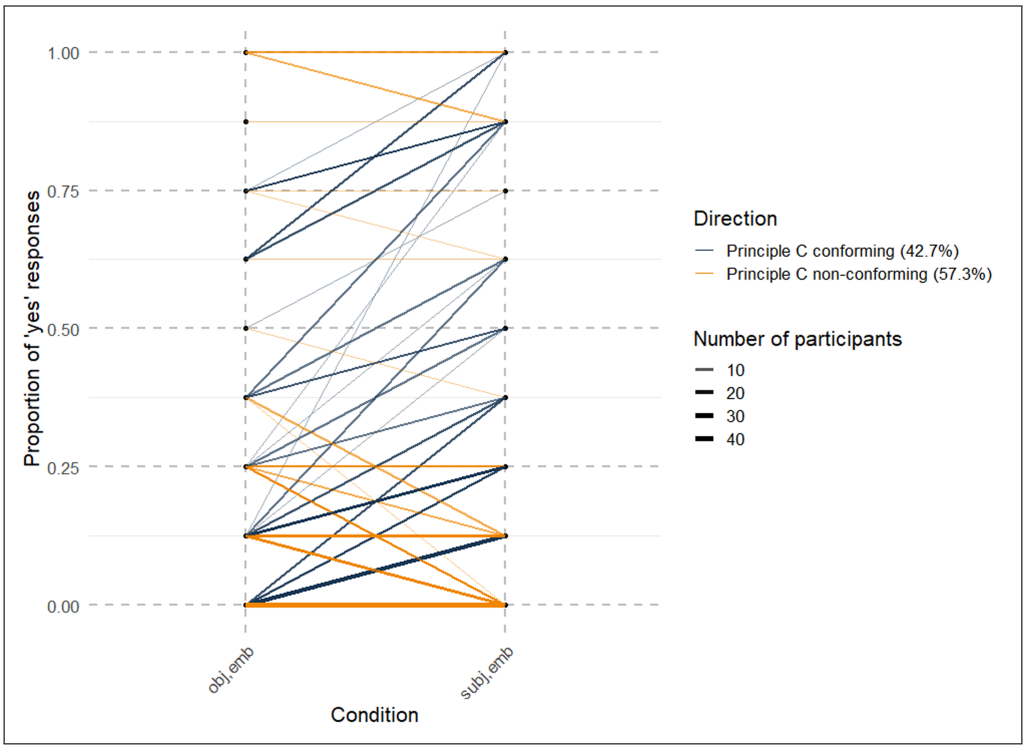

Figure 3: Individual participants' (n = 150) overall proportions of *yes* responses in experiment 1 by condition. Points on y-axis indicate respective proportion (ranging from 0 to 1 in steps of 0.125 based on eight observations per participant per condition), lines connect individual participants' proportions in the condition 'obj,emb' to 'subj,emb'. Color indicates whether the direction of the effect conforms to the predictions based on a principle C violation, thickness and opacity indicate how frequent the respective combination of proportions from the two conditions is in the data set.

Figure 3 visualizes the crucial information regarding how individual participants' responses change from one condition to the other. Only 42.7% of participants show an effect in the predicted direction, while 57.3% show the opposite or no effect at all. Even among the 42.7%, the increase from object to subject condition is nowhere near the predictions of a syntactic constraint. There are barely cases moving from below chance to above chance, let alone indicating the clear possibility of the coreferent reading in the subject condition and the opposite in the object condition. The directionality in the data along with the visual cues is in line with the statistical models, or rather, the problem of fitting any. The data was analyzed in R (R Core Team 2024), modeled by a generalized linear mixed effects model using the *glmer* function with the family *binomial* (logit link) and the optimizer *bobyqa* (Bates et al. 2015). The model only included the conditions asking about the embedded referent, since the conditions inquiring about the matrix referent solely served as a sanity check. The dataset was thus reduced to the two levels of PHRASE as the conditions 'object' and 'subject'. A conservative α-level of 0.05 was defined. The model

included a random effects structure with varying intercepts and slopes for condition, for both participants and items (Barr et al. 2013). Factors were treatment contrast coded such that the intercept is the estimate of the baseline probability of *yes*-responses in the 'subject' condition. The estimated slope indicates the change in log-odds from 'subject' to 'object' condition.

(13) rating ~ phrase + (1 + phrase | item) + (1 + phrase | participant)

This maximal model revealed a perfect negative correlation of the random intercepts and slopes: the higher the intercept of a participant, the lower their slope. Because the model needs to estimate two values per participant and these values have limited informativity due to the correlation, it is likely overfitting and therefore not providing reliable estimates for the fixed effects. Two more models were fitted. One excluded the correlation between the random effects, the other one reduced them to include estimates for random intercepts only. A Likelihood ratio test using the *anova* function revealed that the more complex model, despite its fitting problems, is superior to the models with a simplified random effects structure, indicated by a lower AIC and a significant p-value. Still, the estimates suggest that the maximal model is compensating for noise rather than identifying stable and interpretable effects. This is particularly evident from inspecting the log-transformed estimates and confidence intervals, which reveal that the model is estimating proportion of *yes*-responses to decrease from object to subject condition, and it likewise overestimates the intercepts. The log-transformed estimates and confidence intervals of the model including varying intercepts and slopes are reported below using the function *ggemmeans* (Lüdecke 2018), followed by the summary of the estimates from the maximal model in log odds in **Table 2**. Significance testing revealed no significant effect of PHRASE. Estimates for the random effects by participants are extremely high. The following discussion includes log-transformed estimates, see **Table 3**, with the estimate on the log scale as shown in the table given in parentheses. The intercept by participant, that is, the baseline response of individual participants, is estimated to vary by 0.9986 (6.60).

Given that we are dealing with proportions, that is, values between 0 and 1, this indicates that participants are estimated to vary along the entire possible range of responses. The corresponding slopes, i.e. the effect of the predictor for individual participants, are estimated to vary by 0.6295 (0.53). On the other hand, the baseline response by item is estimated to vary by 0.6341 (0.55), with slopes estimated to vary by 0.5374 (0.15). The model thus also attributes a fair amount of variability to the items. Notice that the estimates for the random effects, particularly by participant, are much higher than for the fixed effects. To evaluate how dominant the random effects truly are, a fourth generalized linear mixed effects model was fitted, this time completely omitting the fixed effect, estimating only the intercept and the random effects structure:

(14) rating ~ 1 + (1 + phrase | item) + (1 + phrase | participant)

	Model 1
(Intercept)	1.96*** (0.27)
phrase	0.24 (0.16)
AIC	1871.64
BIC	1917.91
Log Likelihood	–927.82
Num. obs.	2400
Num. groups: participant	150
Num. groups: item	32
Var: participant (Intercept)	6.60
Var: participant phrase1	0.53
Cov: participant (Intercept) phrase1	–1.88
Var: item (Intercept)	0.55
Var: item phrase1	0.15
Cov: item (Intercept) phrase1	–0.29

Table 2: Coefficients estimated by the maximal model for experiment 1 before log-transformation.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

	x	**predicted**	**std.error**	**conf.low**	**conf.high**	**group**
1	obj	0.920	0.229	0.880	0.947	1
2	subj	0.852	0.223	0.788	0.899	1

Table 3: Summary of log-transformed intercepts and confidence intervals estimated by the maximal statistical model for experiment 1.

A likelihood ratio test comparing the generalized mixed effects model in (13) to the model in (14) revealed that the latter model, i.e. the one entirely excluding the effect of PHRASE, has the superior fit based on a lower AIC and p-value.[6] The dominance of random effects by participant found in the model is in line with the data visualization in **Figure 3**.

[6] To the best of my knowledge, this is a fairly unusual scenario. Even if the predictor is not improving the model's fit, the superiority of the model with random effects only implies that omitting the predictor leads to a better understanding of the data. In other words, rather than having no effect, the predictor appears to have a negative effect on the model fit.

3.2.7 Discussion

The evidence does not suggest that a principle C violation under reconstruction played a role in determining participants' responses in this experiment. Generally, coreference rates are much lower than in the data collected by Salzmann & Wierzba & Georgi (2023), and in particular very low in the subject condition despite no syntactic violation inhibiting coreference. It may be the case that the presence of a matrix referent has a depressing effect on the coreference rate with the embedded referent. There are two potential reasons: on the one hand, participants may be choosing the embedded referent less frequently because the matrix referent is simply too prominent. It is mentioned first in the sentence, it is the subject, and neither of its positions is c-commanded by the pronoun. On the other hand, it may further be preferred over the embedded referent because it is both structurally and linearly more distant from the pronoun (cf. Adger et al. 2017; Stockwell & Meltzer-Asscher & Sportiche 2021). The overwhelming prominence of the matrix referent is supported by comments provided by 18 participants in the conditions inquiring about the embedded referent, some of which indicate ambiguity, others specifically reinforcing the interpretation regarding the matrix referent, which was not offered in the respective trials.

There may also be methodological reasons why coreference rates are generally low. The simplified task may be a poor fit to address the research question as it is. Since it is formulated as a forced choice task with the response options *yes* or *no*, participants are essentially asked to evaluate the truth value of only one reading of a potentially ambiguous sentence. This is not supported by the observation that the matrix referent is given a positive response in nearly all observations. However, one could argue that participants establish coreference with the matrix referent much more readily based on the aforementioned prominence factors, and simply do not consider the additional, weak possibility of the reading involving the embedded referent to be problematic for the truth value judgment regarding the matrix referent reading. Again, this is supported by participants' comments explicitly reinforcing the matrix referent reading despite being asked about the embedded referent. Like in the experiments by Stockwell & Meltzer-Asscher & Sportiche (2021; 2022), participants may have responded based on preferences rather than possibilities.

Closer inspection of the data distribution among participants reveals that the overall coreference rates are the result of computing the mean over extreme inter-individual variability. This is particularly evident from how different statistical models fit the data. Recall that the maximal model, though indicated to be a better fit than a simplified version by a likelihood ratio test, (i) has severe fitting problems due to a perfect negative correlation for random intercepts and slopes, (ii) overestimates the intercepts in general, and (iii) estimates the predictor to have an effect in the opposite direction than indicated by the overall coreference rates in the observed data. Modeling under omission of the predictor revealed that a model based solely on the random

effects structure fit the data best, indicating that the distribution of data is best explained in terms that completely ignore the information of underlying c-command and therefore, a principle C violation under reconstruction. Due to the additional confounds introduced by the design that are discussed above, the conclusion that principle C does not play a role at all may nevertheless be premature. Recall that participants were only asked about one out of two readings per trial, potentially making the *someone else* response dispreferred because the experimental task was poorly chosen altogether considering the fully intended ambiguity of the items.

To summarize, the experiment did not provide data illustrating the previously reported subject-object contrast more straightforwardly, it achieved quite the contrary. Although this may be indebted to the method itself, it is noteworthy that any potential effects traditionally associated with a principle C violation are apparently quite easily overridden, particularly in the subject condition where coreference is not inhibited by any syntactic factors. Under the assumption that there is no principle C reconstruction, a more plausible scenario would involve chance level results. The evidence so far seems to favor a view where relatively low coreference rates are not determined by an underlying principle C violation, but rather by the presence of a competing referent. This issue is explored in the second experiment, which is outlined in the following section.

3.3 Experiment 2: omission of the matrix referent

3.3.1 Materials

In the second experiment, participants saw an interrogative sentence in each trial. The sentence contained a displaced NP modified by a PP containing the only referent in the clause, followed by a pronoun matching the referent (resembling items by Adger et al. 2017; Stockwell & Meltzer-Asscher & Sportiche 2021; 2022). The study followed a Latin square design.

(15) Object condition
[Welche Geschichte über Hanna] fand sie ___ ärgerlich?
which story-ACC about Hanna-ACC find.PST-3SG she-NOM upsetting
'Which story about Hanna did she find upsetting?'
What is this about?
☐ Hanna found a story upsetting. ☐ Someone else found a story upsetting.

(16) Subject condition
[Welche Geschichte über Hanna] ___ hat sie verärgert?
which story-NOM about Hanna-ACC have.PST.3SG she-ACC upset
'Which story about Hanna has upset her?'
What is this about?
☐ A story upset Hanna. ☐ A story upset someone else.

As in experiment 1, 24 target items exploring a different research question were also presented. 12 unambiguous fillers were included, involving a referent and a morphosyntactically mismatched pronoun, making only the response *someone else* acceptable. Participants were guided through three training trials illustrating the task.

3.3.2 Method

Participants were again given a single forced choice task per trial. This time, they were asked what the presented sentence was about and had to decide between two readings. The task thus resembles the task in the experiments by Stockwell & Meltzer-Asscher & Sportiche (2021; 2022). Crucially, instead of two scales judging the naturalness of the offered readings, it was a simple forced choice task measuring preference between the only referent in the sentence and *someone else*.[7] Again, metalinguistic terms were avoided by providing full sentences with the intended reading. The order in which the response options were shown was pseudo-randomized across trials. Sentences were not accompanied by individual context. Instead, participants were given the instruction that they should imagine being at a party and picking up snippets of conversations (Stockwell & Meltzer-Asscher & Sportiche 2021; 2022). Items only appeared in two conditions based on the grammatical function of the displaced wh-phrase, i.e. 'subject' and 'object'. Like in experiment 1, both the surface and the base position of the extracted subject c-commands the object pronoun, and, if there is reconstruction, a principle C violation should only occur in the object condition.

3.3.3 Participants

A total of 64 participants have been recruited over the platform Prolific, with mean age 32.2, with n = 39 identifying as male, n = 24 identifying as female, and n = 1 identifying as non-binary. Monetary compensation was provided to everyone who successfully completed the experiment. Participants fulfilled the same criteria as in experiment 1 (4.69% from Austria, 92.19% from Germany, 3.12% from Switzerland). It was made sure that none of the participants from experiment 1 took part in experiment 2. Although testing different methods with the same participant pool is beneficial to reduce variability, due to data collection for both experiments taking place within a short time window, the priority was to avoid familiarization or adaptation effects from the first experiment affecting the results of the second experiment. Fillers contained a referent and a pronoun with mismatched ϕ-features. Participants who chose the reading indicating coreference between the two in at least 25% of the respective trials were excluded, leaving 60 participants in the sample.

[7] Forced choice was used as an alternative given concerns by Stockwell & Meltzer-Asscher & Sportiche (2021; 2022) about participants not using the scales as they were intended.

3.3.4 Procedure

The experiment was set up using the platform L-Rex (Starschenko & Wierzba 2024). The instructions were the same as in experiment 1. The sentences were shown one by one, with the question *Worum geht es hier?* 'What is this about?' being shown below them. The answer options were repetitions of the event described in the sentence, as shown in (15) and (16).[8] The first block included the training items in non-randomized order, showing two cases where the referent in the sentence could corefer with the pronoun and one case where it could not due to a morphosyntactic mismatch. The second block contained target items, pseudofillers and fillers in pseudo-randomized order, such that two items from the same set of materials were never shown consecutively.

3.3.5 Hypotheses

This experiment measures preferences in contrast to the experiment by Salzmann & Wierzba & Georgi (2023) and experiment 1 reported herein. This has important implications for the expected response rates. If the PP modifier containing the referent reconstructs, coreferent readings should be dispreferred in the object condition, with the embedded referent being chosen well below chance. In the subject condition, chance level performance is expected due to the lack of a syntactic violation. If there is no reconstruction, chance level performance is expected in the object condition, too. Because it is a single forced choice task between two referents, if the proportion to which the embedded referent is chosen increases, the proportion to which the unnamed referent is chosen automatically decreases. Chance level performance indicates no preference for either of the readings. Generally, given the lack of an alternative, higher proportions indicating coreference with the embedded referent are expected. If an increase is observed only in the subject condition, this would indicate an underlying principle C violation. if, however, participants choose the embedded referent above chance across conditions, this would indicate a general bias to resolve pronominal reference with any matching antecedent given in the discourse.

3.3.6 Results

Proportions of responses indicating coreference with the embedded referent across conditions are illustrated in **Figure 4**. Participants chose the answer indicating coreference with the embedded referent in 40.3% of the observations in the object condition. In the subject condition, the embedded referent was chosen in 54% of the observations, indicating no clear preference. There is no matrix referent in the sentence, meaning coreference with it cannot be assessed as a sanity

[8] A reviewer asks whether participants were informed about the ambiguity of some sentences, expressing concerns about participants choosing the *someone else* response over the embedded referent due to being certain about its validity. While participants indeed were not explicitly informed about some sentences being ambiguous, they were told that the study was about preferences. Although the reviewer's concerns are valid, we shall see that they do not seem particularly relevant based on the results of the study.

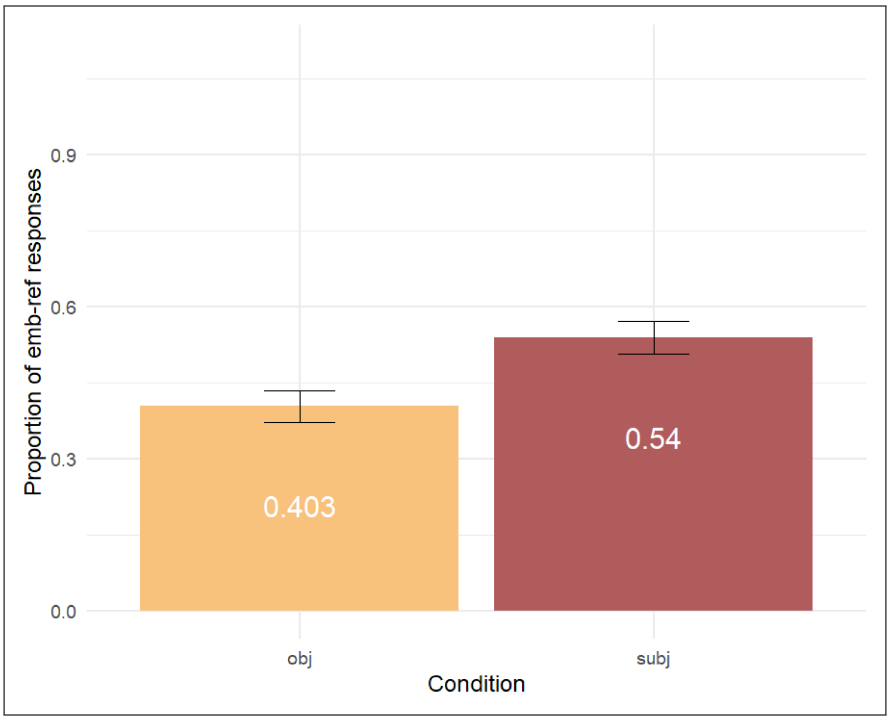

Figure 4: Coreference rates across conditions in experiment 2. Error bars indicate standard error.

check. Participants' performance in the subject condition is in line with the lack of a principle C violation. Confronted with the two possible readings, participants show no clear preference for the embedded referent or the unnamed referent (*someone else*). In the object condition, participants show a slight preference for the option *someone else*, though this is not as strong as previously hypothesized.

Figure 5 shows the distribution of data among participants. Given the omission of the matrix referent, participants each saw 16 items per condition, meaning that the individual proportion of responses indicating coreference with the embedded referent range from 0 to 1, increasing by 0.0625 at a time.

Based on **Figure 5**, it is clear that participants show a vast degree of variability. There is no clustering among participants around chance level. Much rather, participants seem drawn toward the extremes. The majority of participants exhibit combinations of response proportions that are in line with the predictions of a principle C violation. However, pronounced slopes indicating a systematic preference for coreference in the subject but not the object condition and thus a syntactic violation, are extremely rare. 33.3% of participants show no effect at all or a reverse effect.

The data was analyzed in R (R Core Team 2024) using a generalized linear mixed effects model *glmer* with the family *binomial* (logit link) and the optimizer *bobyqa* (Bates et al. 2015)

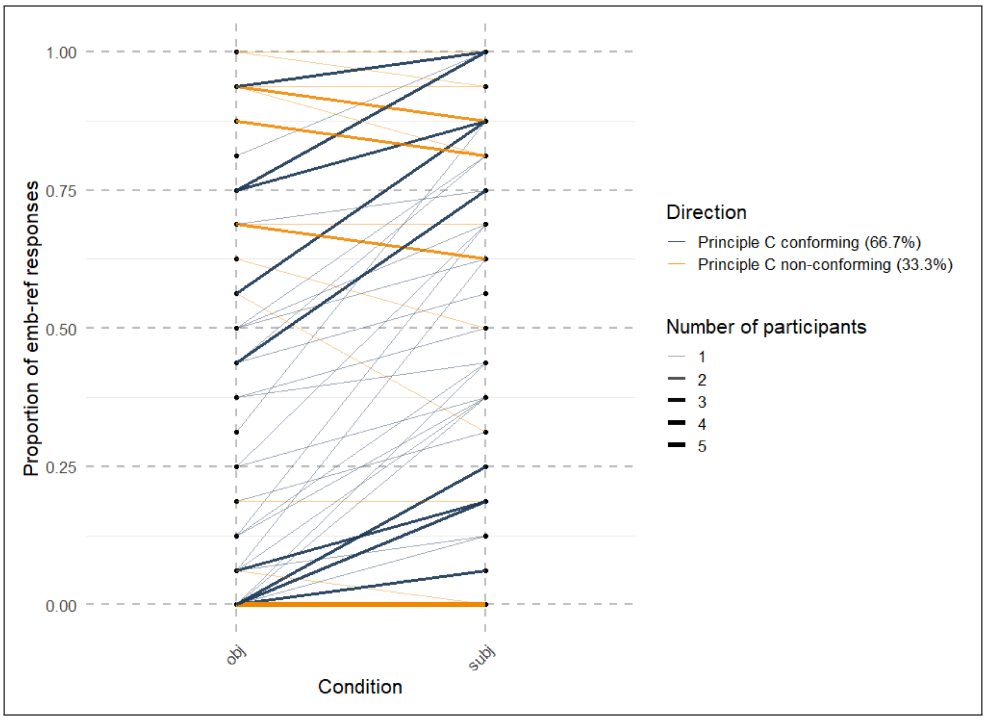

Figure 5: Individual participants' (n = 60) overall proportions of responses indicating coreference with the embedded referent in experiment 2 by condition. Points on y-axis indicate respective proportion (ranging from 0 to 1 in steps of 0.0625 based on 16 observations per participant per condition), lines connect individual participants' proportions in the condition 'obj' to 'subj'. Color indicates whether the direction of the effect conforms to the predictions based on a principle C violation, thickness and opacity indicate how frequent the respective combination of proportions from the two conditions is in the data set.

and a conservative α-level of 0.05. The model included a random effects structure with varying intercepts and slopes both per participants and items. The estimates from the maximal model specified in (17) are reported in **Table 4**.

(17) rating ~ phrase + (1 + phrase | item) + (1 + phrase | participant)

Significance testing revealed a significant effect of PHRASE. As with the data from experiment 1, this model too had fitting problems, revealing a perfect negative correlation of random slopes and intercepts by item, but not by participant. The estimates for the random effects are quite large, i.e. the intercept by participant is estimated to vary by 0.9923 (4.86), while the accompanying slopes are estimated to vary by 0.6735 (0.72). By item, the intercept is estimated to vary by 0.6433 (0.59) and the slopes by 0.5049 (0.02). Model comparison with a simplified model excluding random slopes revealed that the more complex model was nevertheless superior based on a lower AIC and a significant p-value. However, this was again accompanied by large confidence intervals, a slope that is estimated to be much larger than the observed data, and a sign reversal.

	Model 1
(Intercept)	–0.20 (0.33)
phrase	1.19*** (0.18)
AIC	1768.23
BIC	1812.71
Log Likelihood	–876.12
Num. obs.	1920
Num. groups: participant	60
Num. groups: item	32
Var: participant (Intercept)	4.86
Var: participant phrase	0.72
Cov: participant (Intercept) phrase	1.05
Var: item (Intercept)	0.59
Var: item phrase	0.02
Cov: item (Intercept) phrase	–0.11

Table 4: Coefficients estimated by the maximal model for experiment 2 before log-transformation.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

	x	**predicted**	**std.error**	**conf.low**	**conf.high**	**group**
1	subj	0.449	0.330	0.299	0.609	1
2	obj	0.727	0.393	0.553	0.852	1

Table 5: Summary of log-transformed intercepts and confidence intervals estimated by the maximal statistical model for experiment 2.

Upon suspicion that the model was relying on the random effects rather than the the fixed effect to derive the estimates, another generalized linear mixed effects model was fitted omitting the predictor PHRASE altogether (identical to the formula in (14)). However, a likelihood ratio test revealed that the model including the predictor was superior to the random effects only model (p = 1.316e-08), indicating that PHRASE is informative to the model and improves its fit.

The significance test using the maximal model, though possibly inflated, is therefore not entirely misleading. The log-transformed estimates from the maximal model including confidence intervals are summarized in **Table 5**.

3.3.7 Discussion

In this experiment, coreference rates increased compared to the previous studies reported in section 3.1 and 3.2, respectively. This increase is particularly dramatic because the experimental task requires a different interpretation of the results. While in the previous experiment, participants were given a forced choice task between *yes* and *no* concerning one interpretation of a sentence, experiment 2 forced participants to choose between two possible interpretations. In other words, while in prior experiments each referent could score a 100% coreference rate independently (within a single trial in the study by Salzmann & Wierzba & Georgi 2023 and across trials in experiment 1 reported herein), in experiment 2, a 100% choice rate of one referent automatically leads to a 0% choice rate of the other. Participants are at chance level in the subject condition. There is a statistically significant effect leading to 14% fewer responses indicating coreference with the embedded referent in the object condition compared to the subject condition. This shows that a principle C violation decreases the proportion of responses indicating coreference under reconstruction. The overall coreference rate of 40.3% also indicates that isolating c-command as an experimental factor is crucial, since a violation does not rule out the reading completely. **Table 6** summarizes the coreference rates of experiments 1 and 2.

	obj (emb)	**subj (emb)**
Experiment 1	19.4%	26.4%
Experiment 2	40.3%	54%

Table 6: Overall proportion of responses indicating coreference by condition in experiments 1 and 2.

The simplification of the task was an attempt to approximate the setup used by Stockwell & Meltzer-Asscher & Sportiche (2021; 2022). The alternative referent, which intended to ease resolution of pronominal reference in prior experiments, was omitted due to the hypothesis that its presence was distracting from the embedded referent. This turned out to be correct, participants were much more willing to choose the embedded referent in this experiment, where they were given no salient alternative. This is indicative of the importance of non-syntactic factors, such as a trivial bias to assign reference to a referent already present in the discourse (cf. Gordon & Hendrick 1998). The effect of PHRASE is significant, indicating that participants' behavior is at least partially informed by the presence or absence of an underlying principle C violation. The model indicating the significance of this effect nevertheless had severe fitting problems, manifesting in overestimation, large random effects, and a sign reversal. Further inspection of variability through visualizing individual participants' proportions of responses by condition, see **Figure 5**, confirmed that while there is a tendency to respond according to principle C,

participants are generally scattered all across the range of possible responses in both conditions. Inspecting individual participants' data is imperative for understanding the implications of this experiment. It is not the case that participants are generally behaving close to chance level, or that there is a clear clustering. The patterns observed in these data are not merely noisy, they reflect that coreference resolution is guided by a number of factors. A principle C violation under reconstruction, while appearing to be one of them based on experiment 2, appears to play a variable role from participant to participant.

In sum, the experiment yielded a slightly more pronounced contrast between subjects and objects and a significant effect, providing evidence in favor of principle C reconstruction. This is further supported by individual participants' slopes that show an increase in responses indicating coreference from object to subject condition. Consider that the magnitude of this effect does not compare to previous hypotheses about an underlying principle C violation ruling out coreference entirely. While the coreferent reading is chosen less given an underlying principle C violation than in the absence of such a violation, we observe a general bias to resolve coreference with whatever referent is given in the sentence. The principle C reconstruction effect does not hold universally across participants either. The subsequent section places these findings in a broader context.

4 General discussion

4.1 Methodological comparison

We start with a discussion of the impact different methods have on the outcome, which is the key issue I aimed to address with the two novel experiments. In the original study by Salzmann & Wierzba & Georgi (2023), the authors used two forced choice tasks per trial assessing the possibility of either intended reading of the sentence. There were two R-expressions matching the pronoun, one of which was expected to be in disjoint reference with the pronoun in the object condition, given a principle C violation under reconstruction. The authors report a significant effect of PHRASE, indicating that reconstruction has a negative effect on the coreference possibility between the referent contained in the wh-extracted phrase and the pronoun in the object condition, as predicted by principle C. The authors conclude that principle C is a violable constraint in German, and that there is evidence for reconstruction both for adjunct and argument PPs. However, coreference in conditions without the relevant c-command relation in the underlying structure likewise failed to elicit maximal coreference rates. In experiment 1 reported herein, the experimental task was simplified to include only one possible reading, aiming to minimize participants' cognitive load and to facilitate the consideration of both intended readings. The simplifications were hypothesized to yield a more clear-cut result. Instead of increasing coreference rates, this simplification further depressed them overall. The statistical model, which was revealed to be a poor fit for the data under the inclusion of the predictor PHRASE, did not reveal a significant effect. While these findings are clearly in conflict with the

reports by Salzmann & Wierzba & Georgi (2023), the simplification of the task introduced a further confound due to asking participants to judge a truth value for a single reading of an ambiguous sentence, thus not being a suitable task for the provided item structure. Experiment 2 saw further simplifications by omitting the matrix referent and providing participants with a forced choice task between the reading where the pronoun referred to the embedded referent and the reading where it referred to *someone else*, i.e. a referent not present in the discourse.[9] Here, chance level performance is likely indicative of the equal availability of both referents due to assessing the two readings in relation to one another.[10] We observe near chance performance in this experiment, with a significant effect of PHRASE indicating the involvement of principle C in guiding participants' responses. The drastic increase in responses indicating coreference with the embedded referent, however, also suggests that participants prefer to resolve coreference with any referent given in the discourse, even if it is dispreferred in a setup where a prominent alternative is available.

The overall fluctuation in coreference rates across experiments demonstrates how participants' choices are affected by the method that is chosen. This may help understand why the experimental work on English principle C reconstruction led to conflicting conclusions given the diverging methods and setups that were employed (as discussed in section 3.3). The observation that coreference rates can be manipulated via external, non-syntactic factors supports the view that an underlying principle C violation is at best one out of many factors determining coreference in Ā-dependencies. This has far-reaching consequences for the field of theoretical syntax, where principle C reconstruction is widely used for determining c-command relations among constituents as well as testing the properties of movement types, under the assumption that only Ā-movement reconstructs. Considering that such important classifications hinge on coreference judgments often collected from lesser studied languages and using small samples of speakers as well as items, whether the conclusions drawn from such tests tell us anything at all is unlikely in light of the current findings. What is more, as the next section highlights, the conclusions also differ substantially depending on the number and type of speakers one examines.

---

[9] A reviewer objects to the claim that experiment 2 is a simpler than experiment 1, arguing that it does not reduce complexity, but rather relocates it from the item to the responses by including an additional referent there in the form of *someone else*. This is a valid perspective since the task itself is indeed not necessarily simpler. However, navigating the experiment as a whole likely is simpler nonetheless. A named referent in the sentence, especially with highly prominent features such as subjecthood and being mentioned first, is likely a greater distraction from the structure we are actually interested in, i.e. the movement dependency, than an unnamed referent among the response options. A consequence of reducing the item structure is that upon reading the sentence, participants do not necessarily think of an alternative referent by default, making the trials ultimately less demanding.

[10] Note, however, that it may also be a result of uncontrolled pragmatic influences impacting individual participants' responses and canceling each other out in the overall proportions. Section 4.2 discusses this possibility in more detail.

4.2 Inter-speaker variability

This section is devoted to the discussion of variability between participants. The statistical analysis revealed that inter-speaker variability contributes extensively to the understanding of the data composition. Crucially, cases of participants showing the predicted significant increase in coreference rate from object to subject condition are sparse. Many of the participants even show entirely flat slopes, indicating no difference at all between the presence or absence of an underlying principle C violation, while some even show the reverse effect (as shown in **Figures 3** and **5**). Experiment 1 revealed that only 42.7% of participants showed an effect, no matter how minor, in the predicted direction, while in experiment 2, 66.7% of participants did.

The data by Salzmann & Wierzba & Georgi (2023) is re-examined in **Figure 6**. Using the same visualization and criteria, it shows that despite a decrease in overall coreference rates in

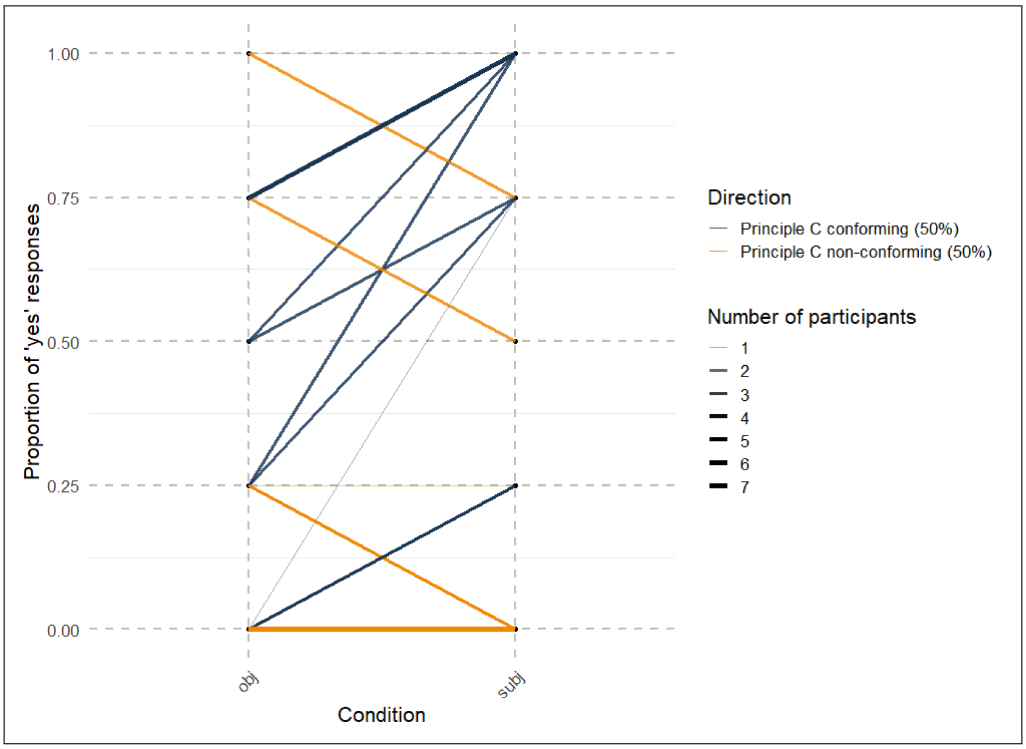

Figure 6: Individual participants' (n = 32) overall proportions of *someone else* responses reported by Salzmann & Wierzba & Georgi (2023) by condition (experiment 2 in the original paper, singling out the conditions 'object, argument, moved' and 'subject, argument, moved' and the task pertaining to the embedded referent). Points on y-axis indicate respective proportion (ranging from 0 to 1 in steps of 0.25 based on 4 observations per participant per condition), lines connect individual participants' proportions from the condition 'obj' to 'subj'. Color indicates whether the direction of the effect conforms to the predictions based on a principle C violation, thickness and opacity indicate how frequent the respective combination of proportions from the two conditions is in the data set.

the presence of a principle C violation, only 50% of participants show an effect in the predicted direction. This measure of directionality is a rather crude test, and yet large portions of the data fail it. Despite its coarseness, the fact that this categorization splits participants into two large groups rather than filtering out a few outliers, provides particularly valuable insights to inter-individual variability. This type of information is entirely lost when simply inspecting overall proportions by condition and whether the predictor is significant in the modeling process. The extremely wide distribution of participants as well as the high percentage of them showing an effect in the opposite direction supports the idea that while an underlying principle C violation plays a role, it is not a particularly strong predictor of coreference resolution in wh-dependencies, and by far not a universal one. If reconstruction were the driving force behind coreference resolution, participants would exhibit a much more pronounced contrast between object and subject conditions, and should, at the very least, by majority show the effect in the predicted direction. The overall spread of participants furthermore indicates the involvement of non-syntactic, but not necessarily non-linguistic factors. They could be prosodic, pragmatic or even linear – and, importantly, have not been controlled for in this experiment, meaning that participants may have constructed prosodic or certain pragmatic cues freely, contributing to the variability in the data. It could also be the case that participants belong to different speaker profiles depending on how relevant principle C reconstruction is in how they resolve coreference. A future longitudinal study could investigate the existence and nature of such speaker profiles – whether they stay stable across methodological and/or pragmatic manipulations and if not, how they vary depending on the type of non-syntactic influence, allowing us to rank the strength of syntactic and non-syntactic cues on coreference resolution. It would be particularly interesting to explore whether certain participants simply do not ever conform to principle C reconstruction, as suggested by the data presented in the experiments herein, and are therefore only sensitive to pragmatic factors. In the next section, I discuss the idea that principle C reconstruction could be optional.

4.3 Optionality of reconstruction

In the following, I touch on what I take the results to indicate, namely that the reconstruction of PP modifiers is optional, including not only inter- but potentially also intra-individual variability based on pragmatic and semantic factors and how these factors are weighted. We saw in section 3.1 that coreference is available to R-expressions in addition to binding and is, according to the more widely accepted view, not determined by syntactic binding principles but by discourse. This observation goes back to the distinction between strict and sloppy readings under ellipsis, showing that R-expressions can bind, but they can also merely refer to the same referent as the pronoun, which does not require any syntactic relation between the two. The standard view is that whenever both operations are available, binding wins over coreference (Grodzinsky &

Reinhart 1993). Notice, however, that the prohibitive nature of principle C requires a rather contrary view. It states that R-expressions need to be free, i.e. they need to *avoid* being bound. A syntactic principle C violation would rule out establishing coreference via means of discourse under the minimalist assumption that pragmatics cannot override syntax.[11] The nature of the phenomenon, however, prompts the question whether reconstruction, if it is optional, can simply be avoided in case it would lead to a syntactic violation. This view would not require us to relativize principle C as such. Late Merger (Lebeaux 1991; 2009) and Neglect (Sportiche 2016) traditionally presuppose an argument-adjunct asymmetry, the relevance of which is quite limited with modifiers of nominals (Salzmann & Wierzba & Georgi 2023 for German, Adger et al. 2017 vs. Stockwell & Meltzer-Asscher & Sportiche 2022 for English). Principle C and even PPs in general are not the problem, however: there is experimental evidence not discussed here which suggests that principle C reconstruction is more stable with PPs of predicates (Adger et al. 2017; Salzmann & Wierzba & Georgi 2023 corroborating Heycock 1995).

A radical alternative is to rethink the theory of movement, i.e. by assuming Parallel Merge instead of a change of syntactic positions leaving behind copies or traces (Citko 2005; 2011; Johnson 2018; 2025). Under such a multidominant theory of movement, the phrase occupies its base and final positions simultaneously, and the effects boil down to which position the phrase is interpreted in by the speaker. Reconstruction effects and their optionality essentially come for free, and the approach may have the added benefit of allowing speakers to differ with respect to how much importance they assign to the non-syntactic cues determining which position to interpret the phrase in.[12] For some, avoiding to refer back to an antecedent with a pronoun that immediately follows it may be a priority (Gordon & Grosz & Gilliom 1993; Arnold 2001). Others may be more lenient with respect to this inhibition, assigning more importance to avoiding the underlying principle C violation. Note, however, that it is not only difficult but also unsolicited to deem these ideas anything more than speculative and a potential incentive for future research. We now take a look at the limitations of the experiments reported herein and how future research could proceed to answer the questions that are left open.

---

[11] A reviewer argues that pragmatics can override syntax because even ungrammatical sentences may be assigned an interpretation and get a boost in acceptability. To pinpoint what it means for pragmatics to override syntax, it is crucial to distinguish acceptability from grammaticality. Grammaticality is one of multiple factors influencing the acceptability of a sentence. That is, as the reviewer points out, an ungrammatical sentence can be acceptable because its acceptability is not exclusively determined by its syntactic grammaticality, but also pragmatic and processing related factors. However, these additional factors, while improving the acceptability of the sentence, cannot override its grammaticality status. See Fanselow (2007) for a series of conceptual arguments defending the view that grammaticality is categorical. Under this view, one would need to argue that cases where coreference is acceptable despite an underlying principle C violaton are necessarily grammatical illusions. While the reviewer is absolutely right that these interactions exist, I am uncertain if modeling them as acceptable ungrammaticalities is desirable.

[12] The tricky part for multidominance is linearizing the structure correctly, ensuring that the element is pronounced in the higher position (Gračanin-Yüksek 2013).

4.4 Limitations

This section addresses the limitations of the current study. First of all, the number of participants and therefore the number of observations decreased from experiment 1 to experiment 2. This is visible in **Tables 2** and **4**. The reason behind this was strictly financial. Notice, however, that despite analyzing the data of 90 participants fewer in experiment 2, there are only around 500 datapoints missing compared to experiment 1. This is indebted to the modifications in experimental design. Recall that experiment 2 omitted the factor REFERENT, which in experiment 1 encoded which referent in the sentence the experimental task was about. Experiment 2 hence has a total of 960 observations, while experiment 1 has a total of 1200. In experiment 1, the 32 items were presented in four conditions, meaning each participant saw eight items per condition. In experiment 2, the number of conditions was reduced to two, meaning each participant saw 16 items per condition. This counterbalances the loss of statistical power to some extent. The result of this is a more uniform sample in experiment 2, which is important to note given the relevance of inter-speaker variability. Experiment 2 has more observations *within* subjects, while experiment 1 has more observations *between* subjects. Assuming that participants do not vary randomly across trials but are systematically guided by specific cues left untested here, the data may show a more pronounced effect of reconstruction due to this reduction in inter-individual variability.

Second, the potential confoundedness of the experimental task employed in experiment 1 likely has an impact on the comparison. Although the matrix referent was given a positive response quite consistently across trials and participants, participants may have been unsatisfied agreeing with the interpretation of the sentence involving the embedded referent because it was much less salient than the matrix referent interpretation. The results would likely turn out differently if participants were confronted with a question phrased in the same manner as done by Salzmann & Wierzba & Georgi (2023): *Can the sentence be understood such that…?*, but only for one of the interpretations per trial.

This brings us to the third major limitation, namely that the possibilities regarding experimental tasks and instructions are virtually endless. Ideally, one would conduct the same experiment over an over again manipulating even more minimal factors. One obvious option would be to omit the task inquiring about the matrix referent from the study by Salzmann & Wierzba & Georgi (2023) and keep the phrasing exactly the same. Further options include prompting participants to give a truth value judgment such as in experiment 1, but omitting the matrix referent. One could also omit the context introduced in experiment 1. As a reviewer points out, experiment 2 did not have a context (due to only featuring a single referent), nor did the experiment by Salzmann & Wierzba & Georgi (2023). Although the context was introduced with the intention of increasing the prominence of the embedded referent compared to the matrix referent by establishing both in the discourse before reading the sentence, the effect may have been canceled out in that both referents could have received an equal boost. It may have even

caused the complete opposite of the intended effect and boosted the prominence of the matrix referent even further without having the desired impact on the embedded referent.

The current investigation does not exhaust the full array of possibilities by any means, but rather aims to provide a starting point, providing evidence that experimentation needs to be carried out carefully and with a special focus on non-syntactic factors. This allows us to use variability in the data as a valuable source of information not only to learn about the phenomena themselves, but to likewise improve the methods we use to study them. As outlined in section 4.2, a longitudinal study on principle C reconstruction across different experimental designs and under the control of pragmatic factors might give us a better understanding of what determines inter-individual differences. With this potential for future investigations in mind, the following section concludes.

5 Conclusion

The experiments reported herein demonstrate that the principle C reconstruction effect is detectable under a specific set of experimental circumstances and, particularly, in large sample sizes. Crucially, for the first time ever, we have data that eliminate the uncertainty associated with comparisons across variable lexical items and experimental designs. While one of the experiments revealed a significant effect, between 33.3% and 57.3% of participants show no effect at all or an effect in the opposite direction. The study comparison revealed that coreference rates vary systematically based on the experimental task and the presence of an alternative referent. This is confirmed by the reexamination of previously published conflicting experiments which vary immensely with respect to these properties. While the isolation of c-command as an experimental factor is crucial to ensure the validity of the test, experiment 1 indicates that even then, the effect may fail to show up due to unrelated manipulations. The study allows us to reevaluate the empirical picture with much more nuance: A principle C violation under reconstruction is at most one out of many factors determining coreference resolution in Ā-dependencies. Whether it is a relevant factor at all seems to depend on the individual speaker. It appears that to pragmaticists and psycholinguists, this is no novel empirical contribution. Rather, the study should prompt theoretical syntacticians to critically rethink their current practices in which principle C reconstruction is, based on this study, falsely attributed a high diagnostic value across situations, speakers and typologically unrelated languages.

Abbreviations

1	first person	AUX	auxiliary	PASS	passive	PST	past
3	third person	DAT	dative	PL	plural	PTCP	participle
ACC	accusative	NOM	nominative	POSS	possessor	SG	singular

Data availability

The data, materials and analysis scripts are available at: https://www.doi.org/10.17605/OSF.IO/R5KTP.

Ethics and consent

The respective funding agency's statement on this matter suggests that no ethical review is required for offline questionnaire studies. The research reported herein did not involve the collection of personally identifiable or sensitive participant data and as such does not pose any risk to participants.

Funding information

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287.

Acknowledgements

I am grateful for feedback from the audiences of the DGfS summer school 2024 and NELS 55, as well as the Potsdam Morpho-Syntax Lab for frequent discussions throughout. I thank Kyle Johnson for challenging my view on syntax with his class on 'Opacity and Multidominance' and subsequent meetings during my visit at UMass Amherst in April 2025. I thank the editor, Sol Lago, for her support during the revision process and two anonymous reviewers for their thorough and constructive comments.

Competing interests

The author has no competing interests to declare.

References

Adger, David & Drummond, Alex & Hall, David & van Urk, Coppe. 2017. Is there Condition C reconstruction? In Lamont, Andrew & Tetzloff, Katerina (eds.), *Proceedings of the 47th Annual Meeting of the North East Linguistics Society (NELS 47)*, 21–31. Amherst, MA: GLSA.

Arnold, Jennifer E. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes* 2. 137–162. DOI: https://doi.org/10.1207/S15326950DP3102_02

Barr, Dale J. & Levy, Roger & Scheepers, Christoph & Tily, Harry J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory of Language* 68. 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Barss, Andrew. 1988. Paths, connectivity, and featureless empty categories. *Annali di Ca'Foscari. Rivista della Facoltà di Lingue e Letterature straniere dell'Università di Venezia* 27(4). 247–279.

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Bianchi, Valentina. 1995. *Consequences of antisymmetry for the syntax of headed relative clauses*. Scuola Normale Superiore, Pisa dissertation. https://www.researchgate.net/publication/245772483_Consequences_of_Antisymmetry_for_the_Syntax_of_headed_relative_Clauses.

Bruening, Benjamin. 2014. Precede-and-command revisited. *Language* 90. 342–388. DOI: https://doi.org/10.1353/lan.2014.0037

Bruening, Benjamin. 2021. Generalizing the presuppositional approach to the Binding Conditions. *Syntax* 24. 417–461. DOI: https://doi.org/10.1111/synt.12221

Bruening, Benjamin & Al Khalaf, Eman. 2019. No argument-adjunct asymmetry in reconstruction for Binding Principle C. *Journal of Linguistics* 55. 247–276. DOI: https://doi.org/10.1017/S0022226718000324

Büring, Daniel. 2005. *Binding theory*. Cambridge: Cambridge University Press.

Citko, Barbara. 2005. On the Nature of Merge: External Merge, Internal and Parallel Merge. *Linguistic Inquiry* 36. 475–497. DOI: https://doi.org/10.1162/002438905774464331

Citko, Barbara. 2011. *Symmetry in syntax: Merge, move, and labels* (Cambridge Studies in Linguistics 129). Cambridge, England: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511794278

Cowles, H. Wind & Walenski, Matthew & Kluender, Robert. 2007. Linguistic and cognitive prominence in anaphor resolution: topic contrastive focus and pronouns. *Topoi* 26. 3–18. DOI: https://doi.org/10.1007/s11245-006-9004-6

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

Fanselow, Gisbert. 2007. Carrots – perfect as vegetables, but please not as a main dish. *Theoretical Linguistics* 33(3). 353–367. DOI: https://doi.org/10.1515/TL.2007.023

Fox, Danny. 1999. Reconstruction, binding theory, and the interpretation of chains. *Linguistic Inquiry* 2(30). 157–196. DOI: https://doi.org/10.1162/002438999554020

Freidin, Robert. 1986. Fundamental issues in the theory of Binding. In Lust, Barbara (ed.), *Studies in the Aquisition of Anaphora*, 151–188. Reidel: Dordrecht. DOI: https://doi.org/10.1007/978-94-009-4548-7_4

Gor, Vera. 2020. *Experimental investigations of Principle C at the syntax-pragmatics interface*. Rutgers University dissertation.

Gordon, Peter C. & Grosz, Barbara J. & Gilliom, Laura A. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science* 17(3). 311–347. DOI: https://doi.org/10.1207/s15516709cog1703_1

Gordon, Peter C. & Hendrick, Randall. 1998. The representation and processing of coreference in discourse. *Cognitive Science* 22. 389–424. DOI: https://doi.org/10.1207/s15516709cog2204_1

Gračanin-Yüksek, Martina. 2013. Linearizing multidominance structures. In Biberauer, Theresa & Roberts, Ian (eds.), *Challenges to linearization*, 269–294. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9781614512431.269

Grodzinsky, Yosef & Reinhart, Tanya. 1993. The innateness of binding and coreference. *Linguistic Inquiry* 1(24). 69–101.

Haegeman, Liliane. 1994. *Introduction to government and binding theory*. Oxford/Malden: Blackwell.

Heim, Irene. 1998. Anaphora and semantic interpretation: A reinterpretation of Reinhart's approach. In Sauerland, Uli & Percus, Orin (eds.), *MIT working papers in linguistics 25*, 205–246.

Heim, Irene. 2007. Forks in the road to rule I. In Abdurrahman, Muhammad & Schardl, Anisa & Walkow, Martin (eds.), *Proceedings of the thirty-eigth annual meeting of the North East Linguistic Society (NELS 38)*, 339–358.

Heim, Irene & Kratzer, Angelika. 1998. *Semantics in generative grammar*. Wiley-Blackwell.

Henderson, Brent. 2007. Matching and raising unified. *Lingua* 117. 202–220. DOI: https://doi.org/10.1016/j.lingua.2005.12.002

Heycock, Caroline. 1995. Asymmetries in reconstruction. *Linguistic Inquiry*. 547–570.

Järvikivi, Juhani & van Gompel, Roger P. G. & Hyönä, Jukka & Bertram, Raymond. 2005. Ambiguous pronoun resolution: Contrasting the first-mention and subject-preference accounts. *Psychological Science* 16(4). 260–264. DOI: https://doi.org/10.1111/j.0956-7976.2005.01525.x

Johnson, Kyle. 2018. *A multidominant theory of movement*. Lectures presented at CreteLing, July 17–27.

Johnson, Kyle. 2025. *Opacity and multidominance*. Lectures presented at the University of Massachusetts, Amherst.

Kaiser, Elsie. 2011. Focusing on pronouns: Consequences of subjecthood, pronominalisation, and contrastive focus. *Language and Cognitive Processes* 26(10). 1625–1666. DOI: https://doi.org/10.1080/01690965.2010.523082

Kuno, Susumu. 2004. Empathy and direct discourse perspectives. In Horn, Laurence R. & Ward, Gregory (eds.), *The handbook of pragmatics*, 315–343. Blackwell publishing. DOI: https://doi.org/10.1002/9780470756959.ch14

Lasnik, Howard. 1998. Some Reconstruction Riddles. In Dimitriadis, Alexis & Lee, Hikyoung & Moisse, Christine & Williams, Alexander (eds.), *Proceedings of the 22nd annual Penn Linguistics Colloquium (PLC 22)*, 83–98.

Lebeaux, David. 1988. *Language acquisition and the form of grammar*. Amherst: University of Massachussetts dissertation.

Lebeaux, David. 1991. Relative clauses, licensing, and the nature of the derivation. In Rothstein, Susan & Speas, Margaret (eds.), *Perspectives on phrase structure: Heads and licensing*, 209–239. San Diego, CA: Academic Press. DOI: https://doi.org/10.1163/9789004373198_011

Lebeaux, David. 2009. *Where does binding theory apply?* The MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262012904.001.0001

Lüdecke, Daniel. 2018. Ggeffects: tidy data frames of marginal effects from regression models. *Journal of Open Source Software* 3(26). 772. DOI: https://doi.org/10.21105/joss.00772

Nissenbaum, Jonathan. 2000. *Investigations of covert phrase movement*. Cambridge, MA: MIT dissertation. https://dspace.mit.edu/handle/1721.1/8842.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Reinhart, Tanya. 1983a. *Anaphora and semantic interpretation*. Chicago: University of Chicago Press.

Reinhart, Tanya. 1983b. Coreference and bound anaphora. *Linguistics and Philosophy* 6. 47–88. DOI: https://doi.org/10.1007/BF00868090

Safir, Ken. 1999. Vehicle change and reconstruction in Ā-chains. *Linguistic Inquiry* 30(4). 587–620. DOI: https://doi.org/10.1162/002438999554228

Sag, Ivan A. 1976. *Deletion and logical form*. MIT dissertation. https://dspace.mit.edu/handle/1721.1/16401.

Salzmann, Martin & Wierzba, Marta & Georgi, Doreen. 2023. Condition C in German A'-movement: Tackling challenges in experimental research on reconstruction. *Journal of Linguistics* 59(3). 577–622. DOI: https://doi.org/10.1017/S0022226722000214

Sauerland, Uli. 1998. *The meaning of chains*. MIT dissertation. https://www.leibniz-zas.de/en/research/publications/details/publications/3384-the-meaning-of-chains.

Sportiche, Dominique. 2016. *Neglect*. Unpublished UCLA manuscript.

Sportiche, Dominique. 2017. Reconstruction, binding, scope. In Everaert, Martin & van Riemsdijk, Henk C. (eds.), *The Wiley Blackwell companion to syntax, Second Edition*. Hoboken, NJ: John Wiley & Sons. DOI: https://doi.org/10.1002/9781118358733.wbsyncom002

Sportiche, Dominique. 2019. Somber prospects for Late Merger. *Linguistic Inquiry* 50(2). 416–424. DOI: https://doi.org/10.1162/ling_a_00306

Starschenko, Alexej & Wierzba, Marta. 2024. *L-Rex Linguistic rating experiments [software], version 1.0.3*. GNU General Public License v3.0. https://github.com/2e2a/l-rex/.

Stockwell, Richard & Meltzer-Asscher, Aya & Sportiche, Dominique. 2021. There is reconstruction for Condition C in English questions. In Farinella, Alessa & Hill, Angelica (eds.), *Proceedings of the 51st annual meeting of the North Eastern Linguistic Society (NELS 51)*, 205–214. Amherst, MA: GLSA.

Stockwell, Richard & Meltzer-Asscher, Aya & Sportiche, Dominique. 2022. Experimental evidence for the Condition C argument-adjunct asymmetry in English questions. In Bakay, Özge & Pratley, Breanna & Neu, Eva & Deal, Peyton (eds.), *Proceedings of the 52nd annual meeting of the North Eastern Linguistic Society (NELS 52)*, 145–158. Amherst, MA: GLSA.

Takahashi, Shoichi & Hulsey, Sarah. 2009. Wholesale late merger: Beyond the A/Ā-distinction. *Linguistic Inquiry* 40(3). 387–426. DOI: https://doi.org/10.1162/ling.2009.40.3.387

Temme, Anne & Verhoeven, Elisabeth. 2017. Backward binding as a psych effect: A binding illusion? *Zeitschrift für Sprachwissenschaft* 36(2). 279–308. DOI: https://doi.org/10.1515/zfs-2017-0011

Van Riemsdijk, Henk & Williams, Edwin. 1981. NP-structure. *The Linguistic Review* 1. 171–217. DOI: https://doi.org/10.1515/tlir.1981.1.2.171

Van Urk, Coppe. 2015. *A uniform syntax for phrasal movement. A case study of Dinka Bor*. Cambridge, MA: MIT dissertation.

Varaschin, Giuseppe & Culicover, Peter W. & Winkler, Susanne. 2023. In pursuit of Condition C: (non-)coreference in grammar, discourse and processing. In Konietzko, Andreas & Winkler, Susanne (eds.), *Information structure and discourse in generative grammar: Mechanisms and processes*. Berlin: De Gruyter.