Bilingualism: Language and Cognition

cambridge.org/bil

Discourse-based pronoun resolution in non-native sentence processing

Cecilia Puebla and Claudia Felser

Potsdam Research Institute for Multilingualism, University of Potsdam, Potsdam, Germany

Research Article

The research reported in this paper has been supported by an Alexander-von-Humboldt Professorship awarded to Harald Clahsen (Potsdam Research Institute for Multilingualism) and by the German Research Foundation (DFG) through Grant No. FE 1138/1-1 awarded to Claudia Felser.

Cite this article: Puebla, C., & Felser, C. (2024). Discourse-based pronoun resolution in non-native sentence processing. *Bilingualism: Language and Cognition*, **27**, 557–571. https://doi.org/10.1017/S1366728923000676

Received: 3 October 2022
Revised: 27 August 2023
Accepted: 4 September 2023
First published online: 25 October 2023

Keywords:
anaphor resolution; variable binding; coreference; L2 processing; eye-movement monitoring; German

Corresponding author:
Cecilia Puebla,
E-mail: cecilia.puebla.antunes@uni-potsdam.de

Abstract

Personal pronouns can potentially be resolved in logical syntax by means of variable binding (VB) or at the discourse-representational level through coreference assignment (CR). Previous research suggests that real-time reference resolution is guided more strongly by discourse-level cues in a non-native language (L2) than in a native language (L1). Here we use the VB/CR distinction to further test this hypothesis. Using eye-movement monitoring during reading and a complementary questionnaire task, we compared L1 German and L1 Russian/L2 German speakers' resolution of object pronouns. While both our participant groups ultimately preferred CR over VB interpretations, only the L2 participants showed evidence of favouring a sentence-external CR antecedent from early on during processing. Our L1 group, by contrast, favoured a VB antecedent during processing. The observed L1/L2 processing differences reveal divergent antecedent search strategies, with L2 but not L1 speakers being primarily guided by discourse-level cues during real-time comprehension.

Introduction

Pronouns can be linked to referential antecedents at the discourse-representational level via coreference assignment (CR) but must be linked to quantified antecedents in logical syntax, via variable binding (VB; Grodzinsky & Reinhart, 1993; Reuland, 2001, 2011). Bound readings for pronouns can only be established intra-sententially, as they are assumed to be mediated by c-command, a hierarchical relationship between sentence constituents (Reinhart, 1983). VB normally requires the quantified expression to c-command the pronoun it binds. CR, on the other hand, is not contingent on c-command and thus can also hold across sentence boundaries (e.g., Reuland, 2011). These two types of referential dependency are illustrated in (1), where the pronoun *she* can receive either a VB or a CR reading.

(1) Isabel Coixet won the Goya award for Best Director with 'The secret life of words'. Every starring actress hoped that she would be invited to an interview soon.

In (1), the pronoun *she* can either be linked to the proper name *Isabel Coixet* via CR or to the universally quantified noun phrase (QP) *every starring actress* via VB. Coreference between the pronoun and the QP is not possible because, by virtue of being quantified, *every starring actress* is a non-referential expression. As the QP c-commands the pronoun, a referential dependency between them can be established via VB, such that the interpretation of the pronoun co-varies with that of the QP. In contrast, the sentence-external noun phrase *Isabel Coixet* is a referential expression denoting a specific individual in the discourse representation, and the name and the pronoun are free to corefer inter-sententially.

Investigating the real-time processing of pronouns in configurations such as (1) can provide a window into the nature of the mental representations, memory operations, and parsing routines underlying reference resolution. Current psycholinguistic models assume that anaphoric expressions are resolved via cue-based memory retrieval (as discussed in more detail in the next section), and the formal distinction between grammatical VB and discourse-based CR might be re-interpreted in processing terms as the prioritization of different types of retrieval cue during comprehension. The VB/CR distinction thus provides a good basis for testing the claim that L2 comprehenders over-rely on discourse and pragmatic cues to interpretation in comparison to L1 comprehenders (Clahsen & Felser, 2006, 2018; Cunnings, 2017a, 2017b). A growing body of bilingual processing studies has found differences in the way L1 and L2 speakers resolve anaphoric expressions in real time despite L2 speakers patterning with L1 speakers in offline tasks (for reviews, see e.g., Cunnings, 2017a; Felser, 2016, 2019; Roberts, 2013). L2 speakers are more prone than L1 speakers to retrieving antecedents that are highly accessible in the current discourse representation. In grammatically constrained

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

configurations, this tendency may result in the initial mis-retrieval of discourse-prominent but grammatically inappropriate antecedents (e.g., Felser & Cunnings, 2012; Felser et al., 2009; Kim et al., 2015). Relative to L1 comprehension, L2 comprehenders' increased sensitivity to discourse-level or contextual information has been shown to enhance their sensitivity to changes in the discourse dynamics such as topic shifts during pronoun resolution (Puebla & Felser, 2022), but also to increase processing costs when additional competitor antecedents are available in the extra-sentential discourse (Roberts et al., 2008).

Examining the processing of pronouns during L2 comprehension can contribute to refining current models of sentence comprehension so as to account for potential cross-population differences in anaphor resolution. If L2 speakers are more likely than L1 speakers to use a discourse-based rather than a syntax-based antecedent search strategy during real-time pronoun resolution, we may expect L2 speakers – but not necessarily L1 speakers – to be drawn to CR antecedents in configurations such as (1), at least during initial processing. Trompelt and Felser (2014) have reported preliminary evidence for L2 speakers favouring CR over VB during processing for intra-sentential CR antecedents that were linearly closer to the pronoun than a potential variable binder. Their study is discussed in more detail below. The current study builds and expands upon Trompelt and Felser's study by examining whether L2 speakers also favour CR antecedents over VB antecedents if the two antecedents' relative proximity to the pronoun is reversed and the CR antecedent is located outside the sentence containing the pronoun.

The psycholinguistics of VB and coreference

From a processing perspective, encountering an anaphoric expression during comprehension is assumed to trigger the retrieval of previously encoded antecedent representations from memory by integrating grammatical, semantic, and discourse-level information. Most current models of language comprehension posit that referential dependencies are resolved through a direct-access, cue-based memory retrieval mechanism. In cue-based parsing, antecedent representations are encoded as feature bundles and stored as memory chunks, which are then re-activated based on how well they match with a composite cue (a retrieval probe) assembled at the retrieval site (the pronoun). The retrieval probe consists of a set of feature-based cues carried by the pronoun and information derived from the current syntactic and discourse context. The antecedent representation that best matches the retrieval probe will show the highest activation level and will be retrieved (e.g., Badecker & Straub, 2002; Lewis & Vasishth, 2005; Parker, 2019; Parker et al., 2017; Van Dyke & McElree, 2011). The retrieval outcome is thus informative about the information sources and constraints that are available to comprehenders during processing.

Both bound-variable and coreference dependencies rely on the faithful encoding and successful re-accessing of information that can be directly implemented as features, such as morphosyntactic information (e.g., grammatical number and gender). However, each type of dependency is additionally constrained by other information types whose exact encoding and implementation as retrieval cues in a cue-based architecture remains unclear. Discourse-based CR relies on conceptual knowledge and on cues drawn from the context and current discourse representation (e.g., topichood) that combine putting one antecedent into focus or rendering it more prominent or accessible than other candidates. VB, on the other hand, is crucially based on the relationship between phrase-structure constituents – notably, c-command, a type of information that is difficult to reduce to feature-based cues (see Kush, 2013; Kush et al., 2015, for discussion).

Another open question for cue-based parsing is how to explain interpretation preferences or differences (e.g., between L1 and L2 speakers) in cases of ambiguity between bound and coreference readings. Cue-based approaches typically assume a uni-modular architecture that uses all information available at the pronoun immediately and in parallel to cue the retrieval of an antecedent (e.g., Parker, 2019). However, while VB can be established using the pronoun's grammatical features and c-command as retrieval cues, CR requires not only morphosyntactic agreement but also a search in the discourse representation and access to pragmatic/conceptual knowledge. Finding L1/L2 processing differences in terms of interpretation preferences can be informative about the ways in which syntactic bottom-up information and contextual top-down cues combine at retrieval, particularly of how much weight L1 and L2 comprehenders assign to the cues and constraints supporting one or the other interpretation. While it used to be assumed that different information types are weighted uniformly (Martin & McElree, 2009, 2011), recent proposals argue for systematic variation in cue weighting across individuals and dependency types (Yadav et al., 2022), and some cue-based models assign preferential weighting and temporal priority to syntactic information when navigating memory representations (e.g., Chow et al., 2014; Parker et al., 2017; Van Dyke & McElree, 2011).

For pronouns that could be linked either to a quantified (VB) or to a referential (CR) antecedent as in (1) above, cue-based retrieval models do not necessarily make any specific predictions as to which antecedent type should be preferred. Linguistic approaches to pronominal reference such as the primitives of binding (PoB) framework (e.g., Grodzinsky & Reinhart, 1993; Koornneef, 2008; Koornneef & Reuland, 2016; Reuland, 2001, 2011), on the other hand, predict that VB should be prioritised over coreference assignment. The PoB includes an economy hierarchy of referential dependencies, according to which computing syntactically-mediated relationships requires less processing cost than establishing discourse-based dependencies. In case of an ambiguity as in (1), the PoB economy principle thus predicts that VB should be preferred over CR (e.g., Koornneef, 2008; Koornneef & Reuland, 2016). While the PoB assumes a modular architecture where VB and CR dependencies are computed by means of different algorithms, within a direct-access, cue-based retrieval approach, the prediction that VB should be prioritised might be captured by assigning relatively more weight to grammatical retrieval cues (including c-command) than to discourse-level or conceptual cues (compare e.g., Parker et al., 2017).

The use of time-course sensitive methods such as eye-tracking during reading has proven a fruitful way to investigate the VB/CR distinction experimentally and to explore comprehenders' real-time pronoun resolution preferences. Several studies by Koornneef and colleagues provide reading-time evidence in favour of a privileged status of bound-variable over CR dependencies during the processing of Dutch subject pronouns (e.g., Koornneef, 2008, 2010; Koornneef et al., 2006). However, conflicting results have been reported in more recent studies – for example, by Cunnings et al. (2014), who examined the resolution of subject pronouns in English. In their eye-tracking experiments, participants read sentences such as (2a,b) below that contained a subject pronoun and two potential antecedent candidates: a QP that c-commanded the pronoun (e.g., *every soldier)* and could

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

only serve as variable binder, and an intra-sentential but non c-commanding proper name (e.g., *James*) that could only possibly enter into a coreference relationship with the pronoun. The order of the antecedents was reversed in the two experiments.

(2) a. *Experiment 1*
Every soldier who knew that {James/Helen} was watching was convinced that {he/she} should wave as the parade passed.

b. *Experiment 2*
It looked to {James/Helen} that every soldier was completely convinced that {he/she} should wave as the parade passed.

The authors used a gender-mismatch paradigm (Sturt, 2003), manipulating the (stereo-)typical gender congruence between the pronoun and each of the two antecedents to obtain four experimental conditions. If VB is computed more easily than CR, as has been posited within the PoB framework, readers should retrieve the QP initially, with the proper name only considered later during processing, or not at all. Retrieval of the QP antecedent should be reflected in QP gender effects – that is, reading-time differences at the pronoun region between sentences containing a gender-matching QP and those containing a mismatching one. Gender effects of the name, if any, should emerge in later measures. Cunnings et al.'s results failed to provide support for a privileged status of VB over CR, however. Instead, gender effects emerged for the linearly closest antecedent to the pronoun, regardless of type, and only in later eye-movement measures. The results from a complementary offline questionnaire confirmed participants' preference for the linearly closest antecedent. The authors concluded that VB is not necessarily prioritised during processing (but see Koornneef & Reuland, 2016, for a different interpretation of these results).

Psycholinguistic accounts of anaphor resolution have been proposed primarily considering monolingual data. Even if the retrieval mechanism weighted grammatical information more strongly than discourse-level cues and VB did indeed take precedence over CR in L1 processing, this might not necessarily be the case during L2 comprehension. Successful models of anaphor resolution should be able to account for potential individual or cross-population differences in reference resolution.

Binding and coreference in L2 comprehension

A growing body of research suggests that L2 learners' real-time comprehension of pronominal anaphors is guided more strongly by non-configurational top-down cues such as the information provided in the extra-sentential context, the discourse dynamics, or the relative discourse-level or syntactic prominence (e.g., topic or subjecthood) of potential antecedents (e.g., Felser & Cunnings, 2012; Felser et al., 2009; Kim et al., 2015; Patterson et al., 2014; Puebla & Felser, 2022; Roberts et al., 2008). Although the formal VB/CR distinction provides a useful test case for examining the time-course of grammatical and extra-grammatical information during L1 and L2 anaphora resolution, studies investigating the L2 processing of pronouns that can be linked either to quantified or referential antecedents are still scarce.

Of direct relevance to the present study is an eye-tracking study by Trompelt and Felser (2014), who used stimulus materials equivalent to those in Cunnings et al.'s experiment 1 (see [2] above) to examine the time-course of VB and CR in L1 and L2 German. The authors recorded the eye-movements of German native speakers and L1 Russian-speaking, advanced learners of German while they read sentences such as (3) below to investigate whether readers preferred to link a subject pronoun (e.g., *er* 'he') to a c-commanding QP (e.g., *jeder Maurer* 'every bricklayer') or to a non c-commanding coreference antecedent (e.g., *Georg*) during processing. Three experimental conditions were constructed in which either one or both antecedents matched the pronoun's gender. Note that, unlike in Cunnings et al.'s (2014) study, readers in Trompelt and Felser's study did not have to rely on the QPs' stereotypical gender as QPs in German are overtly marked for grammatical gender.

(3) *Jeder Maurer, der sah, dass {Georg/Sarah} auf*
every bricklayer$_{MASC}$ who$_{MASC}$ saw that {G./S.} on
der Baustelle war, ahnte, dass {er/sie} heute fleißig
the building site was suspected that {he/she} today hard
arbeiten muss.
work must

'Every bricklayer$_{MASC}$ who saw that {Georg/Sarah} was on the building site suspected that {he/she} would have to work hard today.'

The two participant groups' reading-time profiles diverged. The L1 speakers' eye-movement data showed longer reading times at the pronoun for both mismatching QP and mismatching name antecedents relative to the 'baseline' double-match condition, and no significant differences between the two single-match conditions. This indicates that VB and CR antecedents were equally likely to be considered by the L1 speakers. The L2 participants, by contrast, were significantly slowed down by a gender-mismatching proper name (e.g., *Sarah … er*, in [3]), suggesting a preference for CR. The results from a complementary questionnaire task revealed an ultimate preference for the coreference antecedent in both groups. However, considering Cunnings et al.'s finding regarding the role of antecedent recency, it is conceivable that Trompelt and Felser's L2 speakers dispreferred the QP due to the availability of a linearly closer CR antecedent.

Also using eye-tacking during reading, Felser and Drummer (2022) examined pronoun resolution in German sentences such as (4) below that contained a determiner phrase (DP) antecedent in matrix subject position and a QP antecedent that was linearly closer to the pronoun but embedded within a relative clause (hence, in a non c-commanding position).

(4) *{Der König/ Die Königin}, {der/die} {jeden*
{the king/ the queen} who$_{NOM-MASC/FEM}$ every
Gärtner/jede Gärtnerin}
gardener$_{ACC-MASC/FEM}$
kannte, war überzeugt, dass er mehr Bäume pflanzen sollte.
knew was convinced that he more trees plant should

'{The king/queen}, {who$_{MASC/FEM}$} knew {every gardener$_{MASC/FEM}$}, was convinced that he should plant more trees.'

The results from this study also showed different reading-time patterns across groups. L1 German speakers showed evidence of considering the non c-commanding QP antecedent (e.g., *jeden Gärtner* 'every gardener') throughout processing, as reflected in

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

their longer reading times when the QP's gender mismatched the pronoun's relative to when it matched. This was seen in multiple early and later eye-movement measures both at the pronoun and following region. This reading-time pattern suggests that they were able to extend the QP's scope to allow for the pronoun to be bound in logical syntax (compare also Radó et al., 2019). The native Russian-speaking L2 participants, in contrast, barely considered the QP antecedent but were significantly slowed down by the gender mismatch between the pronoun and the c-commanding DP antecedent (e.g., *der König* 'the king'), albeit in later measures. This pattern suggests that the L2 group tried to resolve the pronoun towards the referential matrix subject. As the referential antecedent was also a potential binder, Felser and Drummer's results do not tell us whether the L2 group's preference for the DP antecedent reflects a preference for CR over binding or a preference for syntactically or discourse-prominent antecedents, however.

In sum, there is evidence suggesting that L2 learners tend to establish pronominal reference primarily on the basis of discourse-level information rather than through binding during real-time comprehension (compare also Felser & Cunnings, 2012, for reflexive pronouns). Learners may follow this strategy to compensate for difficulties computing, manipulating, or re-accessing phrase-structure representations in real time (Clahsen & Felser, 2006, 2018; Felser, 2019), or using relational information such as c-command as retrieval cues (Cunnings, 2017b). The current study seeks to further test the hypothesis that L2 pronoun resolution follows a discourse-based antecedent search strategy.

The present study

Building on and extending Trompelt and Felser's (2014) work, we examined whether and when during processing L1 and L2 speakers of German would consider sentence-external coreference antecedents or sentence-internal binding antecedents for pronouns. We compared the reading patterns of a group of native Russian-speaking L2 learners of German to those of a control group of L1 German speakers in an eye-tracking during reading experiment, after which participants also completed a questionnaire task. However, our materials differed from Trompelt and Felser's in several ways. Besides controlling for the two antecedents' syntactic and semantic salience by making both of them role nouns located in subject position, the sentence containing the pronoun was structurally simpler in our study. We moreover separated the experimental manipulation from the area of measurement (the pronoun) by manipulating the gender features of each potential antecedent rather than those of the pronoun, and we consistently used 3rd person masculine pronouns, which are grammatically unambiguous, unlike the feminine pronouns contained in half of Trompelt and Felser's experimental items. Crucially, we reversed the order of the antecedents and placed them in separate sentences, such that the CR antecedent was sentence-external and thus disfavoured by recency compared to the sentence-internal VB antecedent.

Before turning to the presentation of the two tasks, we first provide a brief characterization of the German and Russian pronominal systems and illustrate VB and CR in both languages. We then begin with the presentation of the questionnaire task, which was administered after the eye-tracking experiment.

VB and CR in German and Russian

Russian and German, two inflecting-fusional languages, display a high degree of typological proximity and a comparable organization of their pronominal systems (Gagarina, 2008; Hakimov, 2021). Both languages show a clear distinction between pronouns and reflexives, with analogous syntactic distributions with both being similarly constrained in terms of binding (Avrutin & Wexler, 1992; Bailyn, 2012). German and Russian personal pronouns have distinct forms for singular and plural number, distinguish between three persons, and are case-marked. Russian has a larger inventory of pronominal forms due to its richer case system, although both paradigms demonstrate a substantial degree of suppletion and syncretism. In both languages, third-person singular pronouns differentiate between masculine, feminine, and neuter genders, but in German and especially in Russian, the neuter and masculine forms are largely homomorphic. Personal pronouns in Russian and German must agree with their antecedents in person, gender, and number. Both languages inflect nouns and adjectives within DPs for number, gender, and case, with varying degrees of syncretism. In German, these grammatical categories are additionally expressed in the article, which Russian lacks (Bailyn, 2012; Gagarina, 2008; Hakimov, 2021; Paperno, 2012; Rakhlin et al., 2015).

In accordance with binding Condition B (Chomsky, 1981), pronouns in German and Russian cannot be bound by a clause-mate antecedent but may co-vary with a clause-external quantified antecedent via VB (Asarina, 2005; Avrutin, 1994; Bailyn, 2012; Haider, 2010), as illustrated in (5).

(5) German: *Der Regisseur war sehr berühmt. Jeder*
Russian: *Režisser byl očen' izvesten. Každyj*
the$_{MASC}$ director was very famous every$_{MASC}$
German: *Schauspieler hoffte, dass man ihn bald zum*
Russian: *aktër nadejalsja, čto ego skoro na*
actor$_{MASC}$ hoped that one him$_{ACC}$ soon to
German: *Interview einladen würde.*
Russian: *Interv'ju priglasjat.*
interview invite would

'The director was very famous. Every actor hoped that he would be invited to an interview soon.'

In (5), the underlined object pronoun (*ihn*; *ego* 'him$_{ACC}$') can be syntactically bound by the intra-sentential c-commanding QP (*jeder Schauspieler*; *každyj aktër* 'every actor$_{MASC}$'), or freely corefer in the discourse with the extra-sentential noun phrase (*der Regisseur*; *režisser,* 'the director$_{MASC}$') via CR. The German and Russian universal quantifiers *jed-e(r/n)* and *každ-yj(-aja/-oe)* ('every') are inflected like singular attributive adjectives: They are overtly marked for case and gender and combine with grammatically singular count nouns only. The German and Russian universal quantifiers also share the following properties: (i) They can be used pronominally (e.g., *jeder hoffte*; *každyj nadejalsja*), and (ii) they have distributional properties (Avrutin & Wexler, 1992; Haider, 2010; Kobele & Zimmermann, 2012; Paperno, 2012).

Given that Russian, like German and English, also allows for bound-variable readings of personal pronouns, a preference for CR interpretations in L2 German would be unlikely to reflect a lack of familiarity with VB.

Questionnaire task

We used an offline antecedent decision task to examine participants' ultimate referential interpretations in the absence of any time pressure.

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

Participants

A group of 50 L1 German speakers (12 male; mean age 26.46 years, age range 19-56, SD 7.25) and a group of 50 L1 Russian-speaking L2 learners of German (7 male; mean age 27 years, age range 19-45, SD 5.53) completed the questionnaire. The majority of L1 participants had knowledge of at least two L2s – English being the most spoken L2 – and seven reported having grown up bilingual, with English as their second native language. The L2 speakers' onset age of acquisition (AoA) of German was seven years or above (mean AoA 17 years, range 7-37, SD 7.13) and their mean length of exposure to German was 10 years (range 1-26 years, SD 6.41) at the time of data collection. To gauge our L2 participants' German proficiency we asked them to complete the paper-and-pencil version of the Goethe placement test (courtesy of the Goethe Institute, 2011). Their mean score was 24.3/30 points, which corresponds to the C1 level of the Common European Framework of Reference for Languages (range 16-30 points, B2-C2, SD 3.51).

Participants were mostly students at University of Potsdam, who were recruited via e-mails, through online advertisements, and by flyers distributed in student accommodations. None of the participants was diagnosed with a language disorder at the time of data collection. Participation in the study was rewarded with either eight Euro or course credit.

Materials

Twelve two-sentence stimulus items were created that contained a 3rd person singular masculine direct (*ihn* 'him$_{ACC}$') or indirect object pronoun (*ihm* 'him$_{DAT}$'), and two gender-matching potential antecedents as shown in (6). The first antecedent (e.g., *der Sekretär* 'the secretary$_{MASC}$') was a DP introduced as the subject of the initial sentence. The following sentence contained the second antecedent (e.g., *jeder Kollege* 'every colleague$_{MASC}$'), which consisted of a QP introduced by the universal quantifier *jeder* ('every') in matrix subject position, and the critical object pronoun (e.g., *ihn* 'him$_{ACC}$') embedded within a declarative impersonal clause with the indefinite subject pronoun *man* ('someone', 'one'). The DP was a sentence-external (and hence, non c-commanding) antecedent that could only be linked to the pronoun via CR, whereas the QP could only serve as a VB antecedent for the pronoun.

(6) *Der Sekretär war neu im Büro.*
the secretary$_{MASC}$ was new in office
Jeder Kollege glaubte, dass man ihn nächstes
every colleague$_{MASC}$ believed that one him$_{ACC}$ next
Jahr befördern würde.
year promote would

'The secretary$_{MASC}$ was new in the office. Every colleague$_{MASC}$ believed that he would be promoted next year.'

The double-match stimulus items were mixed and pseudorandomised with 28 fillers. The fillers were also two-sentence texts containing a pronoun and two potential antecedents, but they varied in other aspects such as the distribution and type of the antecedents (e.g., proper names, indefinite descriptions, QPs introduced by assorted quantifiers). The pronoun was either masculine, feminine, or plural, and was unambiguous in that only one of the two antecedents was grammatically possible.

Procedure

The study was conducted with each participant individually in a single session that lasted for about one hour. Participants were tested in a quiet laboratory room in Postdam or Berlin, and their demographic details were collected prior to the session via a web-based questionnaire. At the beginning of the session, participants were informed in German about the testing protocol and asked to give signed consent. The offline task was a written paper-and-pencil questionnaire that was administered after the online experiment. There was only one version of the questionnaire; it consisted of a binary antecedent decision task investigating whether L1 and L2 participants preferentially interpret ambiguous pronouns via syntactically-mediated binding or via the discourse representation. The questionnaire contained 40 items (including the 12 experimental double-match items), in which the critical pronoun was underlined. Following each item, the two potential antecedents were provided as response options. Participants were asked to select one of them as the preferred referent for the pronoun without thinking about their choice for too long. The order of the two response options was reversed in half of the items to prevent participants from developing a response strategy and to encourage them to pay attention to the task. All participants completed the questionnaire within 10 minutes.

Predictions

Participants' offline preference for either antecedent is indicative of which cues generally lead to the ultimate resolution of the pronoun. A higher proportion of choices of the sentence-external DP antecedent (e.g., *der Sekretär* 'the secretary' in [6]) compared to QP choices is expected if participants' ultimate pronoun interpretations are guided primarily by discourse-based or pragmatic information (e.g., feature-match in the discourse representation, conceptual number). Conversely, a preference for resolving the pronoun intra-sententially based on grammatical feature-matching and c-command (possibly facilitated by antecedent recency) should yield a higher proportion of QP choices (e.g., *jeder Kollege* 'every colleague' in [6]). Although in the questionnaire we expect both participant groups to show awareness of the pronoun's referential ambiguity with respect to coreference and bound-variable readings, with L1/L2 differences becoming evident primarily during real-time processing, cross-population differences in cue weighting may also affect participants' antecedent preferences to the point of finding a between-groups difference offline.

Results

Participants' attention to the task was confirmed by checking their responses to the 28 unambiguous fillers. For these items, both participant groups selected the grammatically correct antecedent with high accuracy (mean filler accuracy: 99% [range 93-100%] and 97% [range 83-100%], for the L1 and L2 group, respectively). The distribution of responses for the 12 double-match stimulus items followed a similar pattern across groups, with DP responses accounting for 63.8% (L1 group) and 68% (L2 group) of total referential choices per group. A proportions test confirmed a statistical difference between QP and DP responses within each participant group (L1: $p < .000$, 95% CI [.00, .39]; L2: $p < .000$, 95% CI [.00, .35]), and a between-groups

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

analysis showed no statistical group differences by response ($z = -1.151$, $p = .25$). These results indicate a dispreference for QP antecedents even when these appear linearly closer to the pronoun and suggest an overall ultimate preference for coreference readings.

Eye-tracking experiment

Our eye-movement monitoring task aimed at investigating whether and when VB and/or CR antecedents for German object pronouns are considered during L1 and L2 comprehension.

Participants

The participants for the eye-tracking experiment were the same who completed the questionnaire, but the data from two participants per group were removed due to track loss. We analysed the eye-movement data from 48 L1 (12 male; mean age 26.63 years, age range 19-56, SD 7.32) and 48 L2 speakers of German (7 male; mean age 27 years, age range 19-38, SD 5.71). Vision in all participants was normal or corrected-to-normal.

Materials

A further 12 short two-sentence texts were constructed, like those used in the questionnaire, yielding a total of 24 stimulus items. Using a gender-mismatch paradigm resulted in three experimental conditions in which either one or both antecedents matched in gender with the critical pronoun as shown in (7a-c).

(7) a. Double-match

Der	*Sekretär*	*war*	*neu*	*im*	*Büro.*
the	secretary$_{MASC}$	was	new	in	office
Jeder	*Kollege*	*glaubte,*	*dass*	*man*	*ihn*
every	colleague$_{MASC}$	thought	that	one	him$_{ACC}$
nächstes	*Jahr*	*befördern*	*würde.*		
next	year	promote	would		

b. QP-mismatch

Der	*Sekretär*	*war*	*neu*	*im*	*Büro.*
the	secretary$_{MASC}$	was	new	in	office
Jede	*Kollegin*	*glaubte,*	*dass*	*man*	*ihn*
every	colleague$_{FEM}$	thought	that	one	him$_{ACC}$
nächstes	*Jahr*	*befördern*	*würde.*		
next	year	promote	would		

c. DP-mismatch

Die	*Sekretärin*	*war*	*neu*	*im*	*Büro.*
the	secretary$_{FEM}$	was	new	in	office
Jeder	*Kollege*	*glaubte,*	*dass*	*man*	*ihn*
every	colleague$_{MASC}$	thought	that	one	him$_{ACC}$
nächstes	*Jahr*	*befördern*	*würde.*		
next	year	promote	would		

'{The secretary$_{MASC/FEM}$} was new in the office. {Every colleague$_{MASC/FEM}$} thought that he would be promoted next year.'

The stimuli were distributed over three presentation lists and randomised with 72 fillers, yielding 96 items per list (plus four practice trials). The set of fillers included 12 items structurally parallel to the experimental items but containing feminine, neuter, or plural object pronouns, as well as antecedents preceded by assorted quantifiers. The remaining 60 fillers varied in their syntactic form and were not necessarily composed of two sentences; 12 fillers did not contain any pronouns. Of the fillers containing pronouns, 24 included either a masculine or a feminine object pronoun, and a further 12 contained other pronoun types (e.g., reflexives, possessives, relatives). Half of all stimulus items were followed by a *yes/no* comprehension question, eight of which directly probed the referent of a pronoun. The full set of experimental materials is available at the Open Science Framework (OSF) website (https://osf.io/nj4kv)

Predictions

The gender-mismatch paradigm allows us to detect whether an antecedent was evaluated for dependency building. If an antecedent is considered, then finding it mismatching the pronoun in gender should elicit longer reading times at or after the pronoun region, and/or more looks back to previous regions of text, relative to the double-match condition (7a; compare Cunnings et al., 2014; Trompelt & Felser, 2014). A slowdown in reading times for the QP-mismatch condition (7b) relative to the double-match condition (7a) will be termed QP effect. Conversely, we will refer to DP effects when the DP-mismatch condition (7c) elicits longer reading times compared to the double-match condition (7a). For measures/regions where both QP and DP effects are observed, a significant difference between the two single-match conditions (7b,c) will tell us whether one antecedent was favoured over the other and to what extent.

If VB is prioritised over CR during processing, we expect QP effects and/or longer reading times for the QP-mismatch condition (7b) compared to the DP-mismatch condition (7c). Modular approaches like the PoB (e.g., Grodzinsky & Reinhart, 1993; Reuland, 2011) and cue-based models that assign a stronger weight to grammatical than to discourse-level information (e.g., Parker et al., 2017) would predict QP effects emerging alone or at a point in time before DP effects arise. Given previous findings by Cunnings et al. (2014), it is possible that the likelihood of retrieving the QP is increased in our study due to the QP's linear proximity to the pronoun.

If, on the other hand, the pronoun is resolved preferentially via feature-matching in the discourse representation, as some L2 processing accounts would predict for L2 speakers (Clahsen & Felser, 2006, 2018; Cunnings, 2017a, 2017b), we might expect the opposite pattern, with DP effects emerging alone or before QP effects, and longer reading times for the DP-mismatch condition (7c) compared to the QP-mismatch condition (7b).

Procedure

During the eye-tracking experiment, participants were sitting in an adjustable chair with their chin resting in a chinrest and their forehead leaning against a bar to avoid head movements as much as possible. The stimuli were presented one by one on a computer screen on a white background in black font (Courier New, bold, 23 pt). The experimental items occupied two lines of text, with the pronoun falling in the middle of the second line. The eye-tracking camera (EyeLink 1000, SR Research) was located below the monitor 50 cm away from the participant's eyes. Reading was binocular but only the right eye was recorded. After successful calibration, the experiment began with brief instructions onscreen followed by four practice trials to familiarise the participant with the procedure. Each trial started with a fixation dot to control for drift. To make the text appear,

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

participants were instructed to look at the dot while pressing the 'continue' button on a gamepad and to read the texts silently at their own pace. By pressing the 'continue' button, participants proceeded with the comprehension question, which they answered by pressing the corresponding *yes/no* button on the gamepad, or directly with the next trial for items not followed by a question. To ensure good data quality and avoid track loss, frequent recalibration was performed over the course of the experiment. The stimulus lists were divided into two blocks of 48 items each that allowed participants a short break in between. Without break, the eye-tracking segment took approximately 40-45 minutes.

Once the online task was completed, the questionnaire was administered. Non-native speakers were additionally asked to complete a short vocabulary and gender selection task, as well as the Goethe placement test containing 30 items. The vocabulary checklist contained 138 critical words (e.g., auxiliary verbs, nouns used as antecedents) crucial for understanding the experimental items. The task consisted of marking any unknown words. The gender task contained 20 feminine and masculine role nouns, representative of different categories of gender formation in German, which participants had seen during the study as potential antecedents. For each of these nouns, participants were asked to select the corresponding gender.

Data analysis

Participants' processing patterns were examined by recording their eye-movements while reading the stimulus items. The eye-tracking during reading method resembles natural reading and constitutes a highly time-course sensitive technique that allows for inspection of initial (first-pass) and later (second-pass) reactions to critical regions of text (Rayner, 1998). For the analysis, we focused on two regions centred around the critical pronoun. The pronoun region contained the word preceding the critical pronoun – the impersonal pronoun *man* – plus the critical pronoun itself (e.g., *man ihn* in [7a-c]). The spillover region contained the following two words (e.g., *nächstes Jahr* in [7a-c]). A disruption associated with the resolution of the dependency is assumed to be reflected in longer reading times or increased regressions to previous parts of the text.

For each region of interest we calculated the following continuous reading-time measures. First-pass reading time amounts to the total time spent in a region from first entering it until exiting it in either direction; this measure can index initial referential decisions. Regression-path time refers to the sum of all fixations in a region from first entering it until leaving it to the right; this measure also includes time spent rereading earlier parts of text and, in our experiment, it may indicate difficulties integrating an antecedent with the pronoun. Rereading time is the total fixation time spent in a region during second-pass readings and it was calculated by subtracting first-pass reading times from total reading times; this measure signals the need to reprocess information. Total reading time is a cumulative measure that corresponds to the summed fixation length for a region; it includes time spent in first- and second-pass readings and is associated with changes in global integration processes. Given that this is a composite measure, if an effect is also found in first-pass time or regression-path time but not in later measures, it may reflect early processes. Conversely, if an effect is also found in second-pass measures, such as rereading time, but not in first-pass measures, the effect might result from later processes (Clifton et al., 2007; Liversedge et al., 1998; but cf. Vasishth et al., 2013). Additionally, we examined three binomial measures. Regressions out measures the probability of the eyes leaving a given region to revisit previous text before reading on, and it typically involves only the first-pass reading of the region; increased regressions out of the interest regions may correlate with effects in regression-path times. Regressions in shows whether a region received regressions from later text; increased regressions into the pronoun may indicate the need to confirm or revise a previously assigned interpretation. Finally, rereading probability measures the likelihood of a region being revisited.

Prior to analysis, experimental trials were individually cleaned. Short fixations (< 80 ms) within one degree of visual angle to an adjacent fixation were automatically merged. All other fixations shorter than 80 ms and longer than 1000 ms were removed. Fixations with vertical drift were manually adjusted (L1 group: 1.4%; L2 group: 2.14% of experimental data). Individual fixations falling between regions and clearly not belonging to a run of fixations were removed (L1 group: 0.36%; L2 group: 0.08% of experimental data). Initially skipped regions were treated as missing data and excluded from statistical analyses (skipping rates: 7 and 1.8% [L1 group]; 2.5 and 0.6% [L2 group], for the pronoun and spillover region, respectively). Three individual trials from the L1 data (0.1% of experimental data) and two trials from the L2 data (0.05% of experimental data) were excluded from statistical analysis due to track loss but were included in accuracy calculations to comprehension questions. Trials in which L2 speakers reported unknown vocabulary were excluded from both statistical analyses and accuracy calculations (2.88% of L2 experimental data).

Statistical analyses were conducted in R (R Core Team, 2016) using linear mixed-effects modelling with the package *lme4* (Bates et al., 2015). The continuous measures were log-transformed to satisfy normality assumptions (Vasishth & Nicenboim, 2016) and analysed with linear mixed-effects models; for the binomial measures, mixed-effects logistic regressions were fitted. To examine potential L1/L2 differences in reading patterns, a between-groups analysis was initially performed on the complete dataset for all measures. The models contained the fixed-factors Condition (DM, QP, DP), sum-coded Group (L1, L2), and a Condition by Group interaction. Planned comparisons were run using treatment contrasts with double-match (DM) as the baseline for Condition. For measures and regions where simultaneous QP and DP effects were observed, the conditions were releveled with QP-mismatch (QP) as the baseline to compare the QP- and DP-mismatch (DP) conditions. All models contained the continuous variable Trial index (centred) as covariate to account for habituation effects as the experiment progressed, as well as by-subject and by-item random slopes for Condition as long as convergence was achieved. In cases of non-convergence, the models were simplified by dropping components one by one as suggested by Barr et al. (2013). For the subsequent per-group analyses, the same procedure was followed but the factor Group was not included in the models.

For measures/regions where significant effects were found, we report the effect sizes (ES). For continuous measures, ES estimates were obtained by fitting a linear mixed model on non-transformed reading times (Jäger et al., 2017) and are given in milliseconds (ms), accompanied by their standard errors (SE) and 95% confidence intervals (CI). In addition, we calculated Cohens' *d* following the formula reported in Brysbaert and

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

Stevens (2018). For binominal measures, ES are expressed in terms of odds ratios (OR) and their 95% CIs.

Results

Accuracy calculations to the end-of-trial comprehension questions showed that both participant groups had paid attention to the task and understood the stimuli. Overall accuracy rates were 95% (range 87-100%, SD 0.03) and 91% (range 80-98%, SD 0.05) for the L1 and L2 group, respectively. Below, we first report a summary of the results from the between-groups analysis followed by a detailed description of the results of the per-group analyses.

Between-groups analysis

Our initial omnibus analysis revealed main effects of Group emerging at both interest regions in all four continuous measures and in regressions out of the spillover region, reflecting the fact that the L2 group was generally slower in reading the stimuli. In addition, we observed numerous between-condition effects across measures and interest regions and a significant DP vs. QP by Group interaction in total reading times at the spillover region ($t = 2.361$, $p = .018$; ES = 112 ms, SE = 51 ms, 95% CI [13, 210]; $d = .20$). Preliminary per-groups analyses for this measure and region revealed the opposite pattern of effects across groups (see results below), which indicates different L1/L2 reading profiles. Recall, however, that total reading times is a composite measure that does not provide information about the time-course of processing. Hence, in order to check for potential L1/L2 time-course differences, we went on to perform separate per-group analyses for the remaining measures and regions. The complete model output for our between-groups analysis can be found in the OSF repository.

L1 group

The L1 group's reading times for the four continuous measures and proportions for the three binomial variables are shown in Table 1; Table 2 provides an overview of the statistical outcomes.

Several effects of Condition emerged at the region containing the pronoun, where the double-match condition (7a) was consistently read faster than the other two conditions (7b,c) and the QP-mismatch condition (7b) elicited the longest reading times. Similarly, for the binomial measures, a higher proportion of regressions in and probability of rereading was found for QP-mismatch (7b) relative to the other two conditions (7a,c), with the double-match condition (7a) the easiest to process. Only in regressions out was the numerical pattern different; in this measure, the double-match condition (7a) produced the highest proportion of regressions and the DP-mismatch (7c), the lowest. No statistically reliable effects were detected in this measure, however.

Between-condition effects proved significant for rereading times, total reading times, and regressions in. While DP effects were restricted to total times (ES = 51 ms, SE = 25 ms, 95% CI [3, 99]; $d = .17$), QP effects arose in rereading times (ES = 81 ms, SE = 31 ms, 95% CI [20, 142]; $d = .27$), total times (84 ms, SE = 30 ms, 95% CI [26, 142]; $d = .24$), and regressions in (OR = 1.56, 95% CI [1.09, 2.24]). The DP vs. QP comparison in total times, where both QP and DP effects were found, revealed no statistical difference between the two conditions.

At the spillover region, we obtained significant QP effects in total reading times (ES = 34 ms, SE = 24 ms, 95% CI [-13, 80]; $d = .16$) and rereading probability (OR = 1.42, 95% CI [1.02, 1.97]), replicating the numerical trend for both measures as well as the QP effects observed in total times at the pronoun region. In both measures, increased costs were found for QP-mismatch (7b) relative to the other two conditions (7a,c).

L2 group

Table 3 shows the L2 group's reading times for the four continuous measures and the proportions for the three binomial ones. The statistical outcomes are presented in Table 4.

At the pronoun region, condition differences became visible in three reading-time measures. In all of them, it was the DP-mismatch condition (7c) that elicited the longest reading times numerically. In total times, we found a significant DP effect (ES = 145 ms, SE = 72 ms, 95% CI [4, 287]; $d = .18$). Additionally, in regressions in and rereading probability, we found a multiple-effect pattern, with both QP and DP effects proving significant (regressions in, QP effect: OR = 1.82, 95% CI [1.28, 2.60], DP effect: OR = 1.97, 95% CI [1.39, 2.82]; rereading probability, QP effect: OR = 1.77, 95% CI [1.26, 2.49], DP effect: OR = 1.89, 95% CI [1.35, 2.67]). The DP vs. QP comparison yielded no statistical difference between these conditions, however.

Concerning the spillover region, condition effects arose in three continuous measures and in two binomial variables. The numerical trend observed at the pronoun region was replicated here, where a mismatching DP (7c) led to increased processing costs compared to the other two conditions (7a,b). Statistically significant DP effects emerged in regression-path times (ES = 111 ms, SE = 54 ms, 95% CI [5, 218]; $d = .17$), rereading times (ES = 216 ms, SE = 73 ms, 95% CI [72, 358]; $d = .20$), total reading times (ES = 161 ms, SE = 51 ms, 95% CI [68, 257]; $d = .19$), regressions out (OR = 1.71, 95% CI [1.09, 2.69]) and rereading probability (OR = 1.71, 95% CI [1.22, 2.40]). In this last measure there was also a significant QP effect (OR = 1.47, 95% CI [1.05, 2.05]), but no statistical difference was found between the two single-match conditions.

Summary of results

The questionnaire results showed that both our L1 and L2 participants ultimately preferred to link the pronoun to the DP antecedent in the first sentence. The analysis of the eye-movement data revealed divergent L1/L2 processing patterns, however. While our native speakers were consistently slowed down by a gender-mismatching QP, the L2 learners were primarily distracted by a mismatching DP. For the L1 group, reliable QP effects emerged in second-pass and cumulative measures at both interest regions, whilst significant DP effects were observed only in total reading times at the pronoun region. The L2 group, by contrast, displayed relatively early DP effects, which persisted across several measures in both interest regions. For this group, significant QP effects were only visible in regressions into the pronoun region and in rereading probability at the pronoun and spillover regions, alongside significant DP effects.

Discussion

We used the distinction between syntactically-mediated VB and discourse-based CR to explore potential differences between L1 and L2 comprehenders' pronoun resolution strategies. Building

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

Table 1. Means in milliseconds and proportions for seven eye-movement measures at the pronoun and spillover regions (L1 group).

		PRONOUN REGION			SPILLOVER REGION		
		Mean	SD	SE	Mean	SD	SE
FIRST-PASS READING TIMES	DM	309	160	8	371	197	10
	QP	332	182	10	363	183	9
	DP	326	168	9	373	205	11
REGRESSION-PATH TIMES	DM	357	203	11	481	363	19
	QP	373	225	12	506	378	19
	DP	368	213	11	506	376	19
REREADING TIMES	DM	370	261	21	420	335	26
	QP	465	343	26	446	344	25
	DP	430	308	24	402	289	22
TOTAL READING TIMES	DM	468	282	15	554	355	18
	QP	557	365	19	587	336	17
	DP	521	319	17	555	319	16
REGRESSIONS OUT	DM	.105	.307	.016	.167	.374	.019
	QP	.084	.278	.015	.192	.394	.020
	DP	.083	.276	.015	.207	.406	.021
REGRESSIONS IN	DM	.262	.441	.023	.189	.392	.020
	QP	.334	.472	.025	.187	.390	.020
	DP	.297	.458	.024	.149	.356	.018
REREADING PROBABILITY	DM	.431	.496	.026	.437	.497	.026
	QP	.483	.500	.027	.503	.501	.026
	DP	.454	.499	.027	.451	.498	.026

Note. SD = standard deviation; SE = standard error

on earlier work by Trompelt and Felser (2014), we examined whether and when during processing L1 and L2 speakers of German would try to link a personal pronoun to a sentence-internal VB or sentence-external CR antecedent. Our experimental materials allowed us to address a potential confound in Trompelt and Felser's study, where the preferred CR antecedent was also the linearly closest one to the pronoun. By placing the CR antecedent outside the sentence containing the critical pronoun, we tested the hypothesis that L2 real-time anaphor resolution is more strongly guided by discourse-level information compared to L1 anaphor resolution irrespective of antecedent recency.

Although both our L1 and L2 speakers ultimately preferred coreference interpretations for ambiguous object pronouns, their reading-time profiles were very different. Our L1 comprehenders showed robust evidence of trying to link the pronoun to the VB antecedent during processing and little evidence of considering the extra-sentential CR antecedent. The L2 group, in contrast, was primarily drawn to the CR antecedent, with a gender-mismatching VB antecedent only affecting the likelihood of their rereading the pronoun and spillover regions. These findings indicate that during real-time comprehension, our L1 participants preferentially tried to resolve the pronoun via VB and our L2 participants via discourse-based CR. In what follows we discuss our results in more detail.

VB and antecedent recency in L1 pronoun resolution

Our L1 speakers' results indicate that this group first tried to resolve the pronoun via VB during processing but ultimately settled on a coreference interpretation. Across different reading-time measures and regions, there was a consistent numerical trend for the double-match condition to elicit the shortest reading times and for the QP-mismatch condition the longest. QP gender effects proved statistically reliable in second-pass and cumulative measures but not in early reading-time measures, which suggests that antecedent retrieval tended to be initiated with some delay. The observed timing of effects is consistent with what has been reported in previous eye-movement studies using a similar design (Cunnings et al., 2014; Trompelt & Felser, 2014). We did not find reliable evidence of the CR antecedent being considered during processing except in total reading times at the pronoun region, where gender effects for both antecedents were observed. L1 speakers' evaluation of the CR antecedent may be related to their ultimate preference of a coreference interpretation of the pronoun, although DP gender effects indicative of participants' considering the CR antecedent were absent from the subsequent spillover region.

The finding that our L1 participants tried to link the pronoun to the VB antecedent during processing but ended up favouring the sentence-external CR antecedent indicates that their initial antecedent search targeted a local antecedent whose grammatical

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

Table 2. Statistical outcomes for seven eye-movement measures at the pronoun and spillover regions (L1 group).

	PRONOUN REGION				SPILLOVER REGION			
	Est.	SE	*t(z)* value	*p* value	Est.	SE	*t(z)* value	*p* value
FIRST-PASS READING TIMES								
QP vs DM	0.052	0.031	1.666	.098	−0.012	0.037	−0.309	.761
DP vs DM	0.043	0.031	1.378	.171	0.002	0.035	0.072	.943
REGRESSION-PATH TIMES								
QP vs DM	0.027	0.033	0.820	.416	0.051	0.041	1.244	.227
DP vs DM	0.014	0.031	0.470	.639	0.053	0.040	1.308	.200
REREADING TIMES								
QP vs DM	0.167	0.064	2.615	.009 *	0.044	0.082	0.528	.601
DP vs DM	0.118	0.065	1.805	.072	−0.047	0.080	−0.596	.557
TOTAL READING TIMES								
QP vs DM	0.140	0.043	3.276	.003 *	0.090	0.042	2.148	.044 *
DP vs DM	0.098	0.041	2.396	.021 *	0.027	0.037	0.740	.464
DP vs QP	*−0.041*	*0.050*	*−0.829*	*.415*	---	---	---	---
REGRESSIONS OUT								
QP vs DM	−0.269	0.291	−0.922	.356	0.178	0.199	0.898	.369
DP vs DM	−0.345	0.294	−1.174	.240	0.249	0.197	1.264	.206
REGRESSIONS IN								
QP vs DM	0.445	0.181	2.467	.014 *	−0.023	0.191	−0.123	.902
DP vs DM	0.247	0.183	1.346	.178	−0.288	0.201	−1.436	.151
REREADING PROBABILITY								
QP vs DM	0.304	0.174	1.745	.081	0.347	0.167	2.085	.037 *
DP vs DM	0.179	0.175	1.025	.305	0.118	0.167	0.708	.479

Note. The two single-match conditions were compared against each other only in those measures/regions where DP and QP effects occurred simultaneously. R-code formula used with *lmer* (continuous measures) and *glmer* (binomial measures): response ~ condition + c.(trial index) + (1 + condition | subject) + (1 + condition | item); Est. = estimate; SE = standard error; * $p < .05$

(although not its conceptual) features matched those of the pronoun, whereas extra-grammatical or discourse-level interpretation cues became dominant at later comprehension stages. Recall that the VB antecedent was a non-referential (and conceptually plural) QP, whilst the CR antecedent was a non-quantificational DP denoting a conceptually singular referent. The CR antecedent might thus be considered a pragmatically more appropriate antecedent and a better match in terms of its conceptual number than the VB antecedent. However, it would appear that our L1 participants first tried to resolve the pronoun via binding, likely using the pronoun's grammatical number, its gender, and c-command as retrieval cues. While gender-matching QPs then provided a good fit, gender-mismatching QPs elicited longer reading times and more processing disruption, in the shape of regressive eye-movements, compared to our double-match baseline condition.

VB allows for the pronoun to be resolved in logical syntax without any need for searching the current discourse representation, which led Koornneef and Reuland (2016) to propose that binding is computationally more economical than CR. Our L1 group's eye-movement patterns are consistent with this hypothesis, although in our stimulus materials, the QP antecedent was also the more recent one, which may have increased the attractiveness of the variable-binding option for our L1 speakers. We saw a preference for the CR antecedent in our untimed questionnaire task, however, which shows that a potential processing advantage for VB over CR does not necessarily lead comprehenders to settle on a VB interpretation. Our questionnaire task allowed participants sufficient time to evaluate the suitability of both antecedents considering the preceding context and the two antecedent candidates' conceptual properties, which ultimately led to the CR antecedent winning out over the VB antecedent.

As noted above, we cannot rule out that antecedent recency also played a role in guiding our L1 speakers' antecedent search. Recall that the materials used by Cunnings et al. (2014) contained two sentence-internal antecedents (see example [2]), and their results indicated that native English speakers were preferentially drawn towards either the VB or the CR antecedent depending on which was the more recent one. Cunnings et al.'s (2014) and the present study may not be directly comparable, however, because of differences in the experimental design and because Cunnings et al. used VB antecedents whose evaluation required participants to access non-grammatical information (stereotypical gender match), whilst CR antecedents were highly accessible due to being proper names (Sanford et al., 1988). In our study, the two antecedents were more similar to each other in that both contained role nouns, and the VB antecedent could be evaluated based on its grammatical features only.

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

Table 3. Means in milliseconds and proportions for seven eye-movement measures at the pronoun and spillover regions (L2 group).

		PRONOUN REGION			SPILLOVER REGION		
		Mean	SD	SE	Mean	SD	SE
FIRST-PASS READING TIMES	DM	378	160	8	525	244	13
	QP	379	169	9	511	245	13
	DP	371	172	9	517	277	15
REGRESSION-PATH TIMES	DM	432	412	22	612	367	19
	QP	429	423	22	606	327	17
	DP	399	201	11	710	679	36
REREADING TIMES	DM	580	502	40	719	592	45
	QP	605	519	37	776	612	43
	DP	711	1224	88	867	1005	70
TOTAL READING TIMES	DM	633	475	25	867	582	30
	QP	711	533	28	938	637	33
	DP	763	989	53	1013	907	48
REGRESSIONS OUT	DM	.049	.217	.011	.103	.304	.016
	QP	.056	.230	.012	.125	.332	.017
	DP	.048	.214	.011	.156	.364	.019
REGRESSIONS IN	DM	.236	.425	.022	.208	.407	.021
	QP	.342	.475	.025	.264	.442	.023
	DP	.348	.477	.025	.232	.423	.022
REREADING PROBABILITY	DM	.440	.497	.026	.476	.500	.026
	QP	.549	.498	.026	.550	.498	.026
	DP	.552	.498	.027	.573	.495	.026

Note. SD = standard deviation; SE = standard error

In short, our L1 speakers' results indicate that they first tried to resolve the pronoun sentence-internally via VB but later also considered extra-sentential information to determine the most appropriate antecedent for the pronoun. Our L1 speakers' eye-movement patterns are compatible with models of anaphor resolution that assume a retrieval mechanism that assigns a stronger weighting to syntactic than to discourse-level information during the real-time resolution of linguistic dependencies (e.g., Parker et al., 2017).

Use of discourse-level information in L2 pronoun resolution

Our L2 learners' offline preference for CR was already reflected in their eye-movement record, which was different from that of the L1 group. During real-time comprehension, our L2 readers were primarily affected by the sentence-external antecedent's gender. This indicates that they preferred resolving the pronoun via discourse-based CR rather than via VB, even though the CR antecedent was further away from the pronoun and separated from it by a sentence boundary.

Significant DP effects emerged shortly after the pronoun was first encountered, in regression-path times and regressions out of the spillover region, and persisted across second-pass and cumulative measures in both regions. We also found effects of the QP's gender in some second-pass measures, along with DP effects, indicating consideration of both antecedents. This reading-time profile shows that our L2 learners favoured CR over VB from shortly after encountering the pronoun onwards, considering the QP only during later processing stages.

Our L2 group's general preference for CR over binding confirms and extends Trompelt and Felser's (2014) findings. Our materials were designed to resolve a potential confound in their study, where the preferred CR antecedent was also the linearly closest one to the pronoun. The current results show very clearly that linear proximity does not determine L2 comprehenders' referential decisions, however, and that the initial antecedent search is not limited to the current sentence. This indicates the preferential use of a syntactically unconstrained memory search strategy rather than readers sequentially searching phrase-structure representations for a matching antecedent. The former strategy involves identifying the best-fitting match for the pronoun among the set of current discourse referents. For readers guided by syntactically unconstrained feature-match in the discourse, the CR antecedent is more attractive that the VB antecedent for several reasons. As the first-mentioned antecedent, it is likely to be particularly salient in the discourse representation, since first-mentioned entities provide a context for encoding the rest of the discourse (e.g., Gernsbacher, 1990; MacWhinney, 1977). The CR antecedent moreover provided a better match for the pronoun than the VB antecedent in terms of its conceptual number, and its referential status should have made it easier for the pronoun to link to than to a non-referential quantified expression (Burkhardt, 2005; Carminati et al., 2002).

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

Table 4. Statistical outcomes for seven eye-movement measures at the pronoun and spillover regions (L2 group).

	PRONOUN REGION				SPILLOVER REGION			
	Est.	SE	*t(z)* value	*p* value	Est.	SE	*t(z)* value	*p* value
FIRST-PASS READING TIMES								
QP vs DM	−0.009	0.031	−0.283	.778	−0.034	0.036	−0.942	.351
DP vs DM	−0.034	0.030	−1.120	.264	−0.029	0.038	−0.774	.446
REGRESSION-PATH TIMES								
QP vs DM	−0.010	0.045	−0.225	.824	−0.006	0.035	−0.161	.873
DP vs DM	−0.041	0.039	−1.040	.310	0.088	0.033	2.637	.013 *
REREADING TIMES								
QP vs DM	0.075	0.074	1.014	.317	0.132	0.073	1.821	.071
DP vs DM	0.122	0.071	1.716	.087	0.159	0.077	2.052	.048 *
TOTAL READING TIMES								
QP vs DM	0.091	0.048	1.901	.070	0.064	0.039	1.647	.108
DP vs DM	0.119	0.051	2.343	.029 *	0.123	0.040	3.063	.004 *
REGRESSIONS OUT								
QP vs DM	0.142	0.348	0.407	.684	0.238	0.235	1.015	.310
DP vs DM	0.038	0.360	0.105	.916	0.535	0.228	2.350	.019 *
REGRESSIONS IN								
QP vs DM	0.599	0.179	3.342	.001 *	0.352	0.184	1.913	.056
DP vs DM	0.679	0.180	3.780	.000 *	0.178	0.189	0.942	.346
DP vs QP	*0.080*	*0.171*	*0.471*	*.638*	---	---	---	---
REREADING PROBABILITY								
QP vs DM	0.569	0.173	3.285	.001 *	0.382	0.170	2.252	.024 *
DP vs DM	0.637	0.174	3.669	.000 *	0.534	0.171	3.127	.002 *
DP vs QP	*0.068*	*0.172*	*0.395*	*.693*	*0.152*	*0.171*	*0.890*	*.374*

Note. The two single-match conditions were compared against each other only in those measures/regions where DP and QP effects occurred simultaneously. R-code formula used with *lmer* (continuous measures) and *glmer* (binomial measures): response ~ condition + c.(trial index) + (1 + condition | subject) + (1 + condition | item); Est. = estimate; SE = standard error; * $p < .05$

Summarising, we found strong evidence for our L2 participants resolving the pronoun in favour of a sentence-external referential DP via discourse-based CR, both offline and during real-time comprehension. For our L2 participants, discourse-level factors, including an antecedent's referential properties and conceptual number, played a more important role than its linear proximity to the pronoun. Our eye-movement results are in line with previous findings showing that L2 speakers attempt to resolve pronominal anaphors via CR rather than binding during processing (e.g., Felser & Cunnings, 2012; Trompelt & Felser, 2014), do not consider QP antecedents in configurations where L1 speakers do consider them (Felser & Drummer, 2022), and are more strongly guided than L1 speakers by extra-sentential discourse information (Puebla & Felser, 2022; Roberts et al., 2008).

Theoretical implications

Taken together, our results show that L1 and L2 comprehenders can arrive at the same final interpretation of pronominal anaphors whilst showing different processing profiles (compare also e.g., Felser & Cunnings, 2012; Patterson et al., 2014). The observed L1/L2 differences indicate cross-population differences in the types of information that are prioritised during real-time comprehension, reflecting different antecedent search strategies. Our findings have implications for theoretical approaches to reference resolution and to L2 processing.

Linguistic theory has postulated that pronouns can potentially be resolved via syntactically-mediated binding or through CR (e.g., Grodzinsky & Reinhart, 1993; Reuland, 2011). Whereas binding relies on grammatical computation, coreference relationships can be established without recourse to grammatical representations. Some experimental evidence suggests that binding may show a processing advantage over coreference during L1 comprehension (e.g., Koornneef, 2008 – but cf. Cunnings et al., 2014; Trompelt & Felser, 2014), and the eye-movement results from our L1 group are consistent with this hypothesis. Our L2 participants' eye-movement patterns, in contrast, indicate that this group primarily relied on the CR route to pronoun interpretation. Unlike the L1 readers, who first considered the pronoun as a bound variable during real-time comprehension, our L2 group read the pronoun as referential from the start. The VB antecedent, which matched the pronoun's grammatical but not its conceptual number, was largely ignored. Our results thus provide evidence for the availability of two alternative antecedent search strategies,

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

a grammatically-based and a discourse-based one. Either one or the other may dominate during processing, with the possibility that the former is more likely to dominate during L1 and the latter during L2 comprehension.

If our conclusion that L1 and L2 speakers tend to use different real-time antecedent search strategies is correct, we need to ask why this might be the case. Our non-native participants were proficient users of German who had no difficulty comprehending our stimulus items. As Russian, the native language of our L2 participants, also allows for bound-variable anaphora, a lack of familiarity with this option is unlikely to be the reason for their disfavouring the binding route in German – especially if Koornneef and Reuland (2016) are correct in that binding is easier to compute than coreference. Recall that establishing discourse-based CR dependencies requires access to extra-grammatical representations that rely on world knowledge, conceptual, pragmatic, and contextual information, as well as the ability to keep track of discourse referents and their relative prominence. Binding relationships, in contrast, can be established in (logical) syntax without considering extra-grammatical or top-down information.

Note, however, that establishing binding relationships presupposes that phrase-structure representations are computed fast enough and in sufficient detail, and that c-command relations between constituents are well defined and stable enough to be quickly recovered and implemented as a retrieval cue. Our results suggest that this may be the case in native but not necessarily in non-native anaphor resolution. As L2 speakers have often been reported to under-use grammatical information during real-time processing, our L2 group's discourse-based antecedent search strategy may stem from difficulties encoding, maintaining, or navigating phrase-structure representations, or using c-command as a retrieval cue during real-time comprehension (e.g., Clahsen & Felser, 2018; Cunnings, 2017a, 2017b; Felser, 2016, 2019). Alternatively, L2 speakers may favour CR because they find discourse-level cues generally easier to access and/or to implement for retrieval as compared to other types of cues, including other extra-grammatical cues (Kaiser, 2017). Further research is needed to systematically examine the role of different types of discourse-level and other non-grammatical cues to tease apart these two possibilities.

Compared to native speakers, L2 learners have been found to be more sensitive than L1 speakers to extra-sentential context information during reference resolution (e.g., Puebla & Felser, 2022; Roberts et al., 2008) and tend to favour pragmatically salient and/or discourse-prominent antecedents regardless of their accessibility in terms of binding (e.g., Felser & Cunnings, 2012; Kim et al., 2015; Patterson et al., 2014). The current results are in line with these findings and point towards cross-population differences in the relative weighting of information sources during processing, with L2 comprehension relying more heavily on discourse-level, conceptual, or contextual-pragmatic cues than on structure-based information compared to L1 processing. Our L2 results show that a discourse-based pronoun resolution strategy does not necessarily take more time than a structure-based approach or result in non-nativelike interpretive preferences. The fact that our L2 group integrated the extra-sentential antecedent more readily than their native speaker counterparts suggest that there may be situations in which retrieving an antecedent using a discourse-based search may be more efficient than searching phrase-structure representations. Assuming that these two alternative antecedent search strategies do indeed co-exist, future research might want to explore the role of individual-difference variables other than language status, such as proficiency, AoA or working memory, for determining which strategy is likely to dominate.[1]

Limitations and outlook

The current study provides evidence for L2 speakers, but not L1 speakers, favouring a sentence-external CR antecedent over a sentence-internal VB antecedent throughout processing. This is a novel finding that extends the results from previous studies on L2 anaphora resolution which focused on intra-sentential competitor antecedents. Our study was designed to test the hypothesis that L2 pronoun resolution follows a discourse-based antecedent search strategy, with discourse-level retrieval cues being weighted more strongly compared to L1 processing. While our results support this hypothesis, our design does not allow us to distinguish whether L2 speakers' discourse-based strategy results from difficulties computing or re-accessing syntactic representations, or from discourse-based cues being easier to implement for L2 speakers than other types of cue (Kaiser, 2017). Relatedly, even though our results provide evidence for the co-existence of two alternative antecedent search strategies, it was not our aim to empirically evaluate modular (such as the PoB model) vs. non-modular approaches to anaphor resolution. Assuming that the observed group differences reflect differences in the relative weighting of antecedent retrieval cues during processing, future research is warranted that systematically examines the underlying causes for the observed L1/L2 differences in cue weighting.

Conclusion

While L1 and L2 speakers of German both ultimately settled on a coreference interpretation of object pronouns, they differed substantially in their reading-time profiles. The divergent L1/L2 reading-time patterns show that pronoun resolution can be attempted via two alternative antecedent search strategies, depending on what information sources are available and used by comprehenders as retrieval cues. Our native speakers initially favoured a local c-commanding, non-referential antecedent during processing. Our L2 group, on the other hand, tried to link the critical pronoun to a sentence-external coreference antecedent while essentially ignoring a potential local binder until later processing. The observed cross-population processing differences can be captured by cue-based retrieval models that allow for individual or cross-population differences in the weighting of information sources (Yadav et al., 2022). The current findings provide further evidence in favour of the hypothesis that L2 processing is more sensitive to discourse-level information sources compared to L1 processing (Clahsen & Felser, 2006, 2018; Cunnings, 2017a; Felser, 2019).

Data availability statement. The experimental materials, analyses and data that supports the findings of this study are openly available via the Open Science Framework at https://osf.io/nj4kv

Acknowledgements. We are grateful to our colleagues of the Potsdam Research Institute for Multilingualism for fruitful discussion, the audience at the AMLaP Conference 2021 for their insightful comments, and João Veríssimo for providing statistical advice. We especially thank the participants who kindly volunteered to take part in this study.

Competing interest. The authors declare none.

Publishing ethics. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The current study was approved by the ethics committee of the University of Potsdam (application 44/2017).

Notes

[1] Although the current study was not designed to examine the impact of individual-difference variables on L2 anaphora resolution, we followed a reviewer's suggestion and carried out additional post-hoc analyses of the L2 group to check for potential effects of German proficiency and AoA of German on our experimental conditions. Alongside main effects of both variables across measures and regions, reflecting overall slower reading times with increasing AoA and lower proficiency, we observed a significant DM vs. DP by AoA interaction in first-pass times at the pronoun region (t = -2.051, p = .041). This interaction shows that the later the onset of acquisition of German, the more likely it is for DP effects to emerge during initial processing, which further points to cross-population processing differences in the timing/weighting of discourse-level cues as a function of language status. The complete model outputs of these post-hoc analyses can be found in the OSF repository.

References

Asarina, A. (2005). Russian binding theory: Two improved movement approaches. Unpublished manuscript. MIT. https://web.mit.edu/alya/www/binding.pdf

Avrutin, S. (1994). The structural position of bound variables in Russian. *Linguistic Inquiry*, *25*(4), 709–727. http://www.jstor.org/stable/4178882

Avrutin, S., & Wexler, K. (1992). Development of Principle B in Russian: Coindexation at LF and Coreference. *Language Acquisition*, *2*(4), 259–306. https://doi.org/10.1207/s15327817la0204_2

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 748–769. https://doi.org/10.1037//0278-7393.28.4.748

Bailyn, J. F. (2012). *The syntax of Russian*. Cambridge: Cambridge University Press.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1): 9, 1–20. https://doi.org/10.5334/joc.10

Burkhardt, P. (2005). *The syntax-discourse interface: Representing and interpreting dependency*. Amsterdam & Philadelphia: John Benjamins.

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

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

Chow, W. Y., Lewis, S., & Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in Psychology*, *5*, 630. https://doi.org/10.3389/fpsyg.2014.00630

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42. https://doi.org/10.1017/S0142716406060024

Clahsen, H., & Felser, C. (2018). Some notes on the Shallow Structure Hypothesis. *Studies in Second Language Acquisition*, *40*(3), 693–706. https://doi.org/10.1017/S0272263117000250

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In van Gompel R. P. G., Fischer, M. H., Murray, W. S., & Hill, R. L. (eds.), *Eye movements: A window on mind and brain*. Elsevier, pp. 341–371. https://doi.org/10.1016/B978-008044980-7/50017-3

Cunnings, I. (2017a). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, *20*(4), 659–678. https://doi.org/10.1017/S1366728916000675

Cunnings, I. (2017b). Interference in native and non-native sentence processing. *Bilingualism: Language and Cognition*, *20*(4), 712–721. https://doi.org/10.1017/S1366728916001243

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

Felser, C. (2016). Binding and coreference in non-native language processing. In Holler, A., & Suckow, K. (eds.), *Empirical perspectives on anaphora resolution*. Berlin, Boston: De Gruyter, pp. 241–266. https://doi.org/10.1515/9783110464108

Felser, C. (2019). Structure-sensitive constraints in non-native sentence processing. *Journal of the European Second Language Association*, *3*(1), 12–22. http://doi.org/10.22599/jesla.52

Felser, C., & Cunnings, I. (2012). Processing reflexives in English as a second language: The timing of structural and discourse-level constraints. *Applied Psycholinguistics*, *33*(3), 571–603. https://doi.org/10.1017/S0142716411000488

Felser, C., & Drummer, J. D. (2022). Binding out of relative clauses in native and non-native sentence comprehension. Journal of Psycholinguistic Research, *51*, 763–788. https://doi.org/10.1007/s10936-022-09845-z

Felser, C., Sato, M., & Bertenshaw, N. (2009). The on-line application of binding Principle A in English as a second language. *Bilingualism: Language and Cognition*, *12*(4), 485–502. https://doi.org/10.1017/S1366728909990228

Gagarina, N. (2008). Anaphoric pronominal reference in Russian and German narratives: bilingual and monolingual settings. *Zeitschrift für Slawistik*, *53*(3), 326–338. https://doi.org/10.1524/slaw.2008.0023

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Goethe Institute. (2011). German Placement Test. https://www.goethe.de/de/spr/kup/tsd.html

Grodzinsky, Y., & Reinhart, T. (1993). The innateness of binding and of coreference. *Linguistic Inquiry*, *24*(1), 69–101. https://www.jstor.org/stable/4178802

Haider, H. (2010). *The syntax of German*. New York: Cambridge University Press.

Hakimov, N. (2021). *Explaining Russian-German code-mixing: A usage-based approach*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.5589446

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

Kaiser, E. (2017). On the role of discourse-level information in second-language sentence processing. *Bilingualism: Language and Cognition*, *20*(4), 698–699. https://doi.org/10.1017/S1366728916001012

Kim, E., Montrul, S., & Yoon, J. (2015). The on-line processing of binding principles in second language acquisition: Evidence from eye tracking. *Applied Psycholinguistics*, *36*(6), 1317–1374. https://doi.org/10.1017/S0142716414000307

Kobele, G. M., & Zimmermann, M. (2012). Quantification in German. In Keenan, E. L., & Paperno, D. (eds.), *Handbook of quantifiers in natural language*. Dordrecht: Springer, pp. 227–283.

Koornneef, A. (2008). *Eye-catching anaphora*. Utrecht: LOT International Dissertation Series.

Koornneef, A. (2010). Looking at anaphora: The psychology reality of the primitives of binding model. In Everaert, T., Lentz, H., de Mulder, Ø., Nilsen, A., & Zondervan, A. (eds.), *The linguistic enterprise: From knowledge of language to knowledge of linguistics*. Amsterdam: John Benjamins, pp. 141–166.

Koornneef, A., & Reuland, E. (2016). On the shallow processing (dis)advantage: Grammar and economy. *Frontiers in Psychology*, *7*, 82. https://doi.org/10.3389/fpsyg.2016.00082

Koornneef, A., Wijnen, F., & Reuland, E. (2006). Towards a modular approach to anaphor resolution. In Artstein, R., & Poesio, M. (eds.), *Ambiguity in Anaphora Workshop Proceedings*, Malaga: ESSLI, pp. 65–72.

Kush, D. (2013). Respecting relations: Memory access and antecedent retrieval in incremental sentence processing. Ph.D. dissertation, College Park: University of Maryland. http://hdl.handle.net/1903/14589

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press

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

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

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In Underwood, G. (ed.), *Eye guidance in reading and scene perception*. Elsevier Science Ltd, pp. 55–75. https://doi.org/10.1016/B978-008043361-5/50004-3

MacWhinney, R. (1977). Starting points. *Language*, *53*(1), 152–168. https://doi.org/10.2307/413059

Martin, A., & McElree, B. (2009). Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1231–1239. https://doi.org/10.1037/a0016271

Martin, A., & McElree, B. (2011). Direct-access retrieval during sentence comprehension: Evidence from sluicing. *Journal of Memory and Language*, *64* (4), 327–343. https://doi.org/10.1016/j.jml.2010.12.006

Paperno, D. (2012). Quantification in standard Russian. In Keenan, E. L., & Paperno, D. (eds.), *Handbook of quantifiers in natural language*. Dordrecht: Springer, pp. 729–779.

Parker, D. (2019). Cue combinatorics in memory retrieval for anaphora. *Cognitive Science*, *43*(3), e12715. https://doi.org/10.1111/cogs.12715

Parker, D., Shvartsman, M., & Van Dyke, J. (2017). The cue-based retrieval theory of sentence comprehension: New findings and new challenges. In Escobar, L., Torrens, V., & Parodi, T. (eds.), *Language processing and disorders*. Newcastle: Cambridge Scholars Publishing, pp. 121–144.

Patterson, C., Trompelt, H., & Felser, C. (2014). The online application of binding condition B in native and non-native pronoun resolution. *Frontiers in Psychology*, *5*, 147. https://doi.org/10.3389/fpsyg.2014.00147

Puebla, C., & Felser, C. (2022). Discourse prominence and antecedent misretrieval during native and non-native pronoun resolution. *Discours*, *29*. https://doi.org/10.4000/discours.11720

Radó, J., Konietzko, A., & Sternefeld, W. (2019). Telescoping in relative clauses: Experimental evidence. In Krifka, M., & Schenner, M. (eds.), *Reconstruction effects in relative clauses*. Berlin/Boston: De Gruyter, pp. 405–426. https://doi.org/10.1515/9783050095158-013

Rakhlin, N., Kornilov, S. A., Reich, J., & Grigorenko, E. L. (2015). Interpretation of anaphoric dependencies in Russian-speaking children with and without developmental language disorder. *Language Acquisition*, *22*(4), 355–383. https://doi.org/10.1080/10489223.2015.1028629

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org

Reinhart, T. (1983). *Anaphora and semantic interpretation*. London: Croom Helm.

Reuland, E. (2001). Primitives of binding. *Linguistic Inquiry*, *32*(3), 439–492. https://doi.org/10.1162/002438901750372522

Reuland, E. (2011). *Anaphora and language design*. Cambridge, MA: MIT Press.

Roberts, L. (2013). Sentence processing in bilinguals. In van Gompel, R. P. G. (ed.), *Sentence processing*. Hove: Psychology Press, pp. 221–246.

Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Studies in Second Language Acquisition*, *30*(3), 333–357. https://doi.org/10.1017/S0272263108080480

Sanford, A., Moar, K., & Garrod, S. (1988). Proper names as controllers of discourse focus. *Language and Speech*, *31*(1), 43–56. https://doi.org/10.1177/002383098803100102

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

Trompelt, H., & Felser, C. (2014). Variable binding and coreference in non-native pronoun resolution. In Orman, W., & Valleau, M. J. (eds.), *BUCLD 38: Proceedings of the 38th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, pp. 471–483.

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

Vasishth, S., & Nicenboim, B. (2016). Statistical methods for linguistic research: Foundational ideas – Part I. *Language and Linguistics Compass*, *10*(8), 349–369. https://doi.org/10.1111/lnc3.12201

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *WIREs Cognitive Science*, *4*(2), 125–134. https://doi.org/10.1002/wcs.1209

Yadav, H., Paape, D., Smith, G., Dillon, B., & Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using approximate Bayesian computation. *Open Mind*, *6*, 1–24. https://doi.org/10.1162/opmi_a_00052

https://doi.org/10.1017/S1366728923000676 Published online by Cambridge University Press