Creemers, A., & Meyer, A. S. 2022. The processing of ambiguous pronominal reference is sensitive to depth of processing. *Glossa Psycholinguistics* 1(1): 3, pp. 1–30. DOI: https://doi.org/10.5070/G601166

The processing of ambiguous pronominal reference is sensitive to depth of processing

Ava Creemers, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, Ava.Creemers@mpi.nl

Antje S. Meyer, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands; Radboud University Nijmegen, the Netherlands, Antje.Meyer@mpi.nl

Previous studies on the processing of ambiguous pronominal reference have led to contradictory results: some suggested that ambiguity may hinder processing (Stewart, Holler, & Kidd, 2007), while others showed an ambiguity advantage (Grant, Sloggett, & Dillon, 2020) similar to what has been reported for structural ambiguities. This study provides a conceptual replication of Stewart et al. (2007, Experiment 1), to examine whether the discrepancy in earlier results is caused by the processing depth that participants engage in (cf. Swets, Desmet, Clifton, & Ferreira, 2008). We present the results from a word-by-word self-paced reading experiment with Dutch sentences that contained a personal pronoun in an embedded clause that was either ambiguous or disambiguated through gender features. Depth of processing of the embedded clause was manipulated through offline comprehension questions. The results showed that the difference in reading times for ambiguous versus unambiguous sentences depends on the processing depth: a significant ambiguity penalty was found under deep processing but not under shallow processing. No significant ambiguity advantage was found, regardless of processing depth. This replicates the results in Stewart et al. (2007) using a different methodology and a larger sample size for appropriate statistical power. These findings provide further evidence that ambiguous pronominal reference resolution is a flexible process, such that the way in which ambiguous sentences are processed depends on the depth of processing of the relevant information. Theoretical and methodological implications of these findings are discussed.

Introduction

Recent work by Grant et al. (2020) made explicit a discrepancy in results on the processing of *structural* ambiguities, in which the syntactic structure is ambiguous, and *referential* ambiguities, in which the reference of an anaphoric expression is ambiguous. While some results indicate that ambiguous sentences are read faster than unambiguous ones (i.e., an ambiguity *advantage*), others show that ambiguous sentences slow down reading (i.e., an ambiguity *penalty*). In particular, studies of structural ambiguities typically report that ambiguity facilitates processing (e.g. Traxler, Pickering, & Clifton, 1998; Van Gompel, Pickering, Pearson, & Liversedge, 2005; Van Gompel, Pickering, & Traxler, 2001, for an overview see Clifton and Staub 2008). In contrast, studies on the processing of ambiguous pronominal reference typically suggest that ambiguity may hinder referential processing (Badecker & Straub, 2002; Stewart et al., 2007).

To examine the difference in the processing of structural and referential ambiguities, Grant et al. (2020) directly compared the two ambiguity types and argued that they are processed similarly when assessed under the same conditions. The results showed an ambiguity advantage for the processing of sentences that contained an ambiguous pronominal reference (e.g., *We met the brother of the waiter when* ***he*** *visited the restaurant*). A similar ambiguity advantage was found for sentences that involve a structural ambiguity in the form of a prepositional phrase attachment ambiguity (e.g., *We met the brother of the waiter* ***with a beard***). These findings prompted an investigation of the cause of the discrepancy between studies that find an ambiguity advantage versus studies that find an ambiguity penalty.

There is some evidence based on the processing of *structural* ambiguities suggesting that comprehenders may underspecify certain linguistic relations when they only engage in shallow, rather than deep processing (Swets et al., 2008, see also Logačev and Vasishth 2016b). If, as Grant et al. (2020) suggest, the language processing system handles the resolution of attachment ambiguities and pronominal reference ambiguities in a similar way, we expect that the processing of sentences with *ambiguous pronouns* should also be sensitive to processing depth. As will be discussed in more detail, a study by Stewart et al. (2007) provided initial empirical support for the influence of depth of processing on the processing of ambiguous pronominal reference. In the present study, we provide a conceptual replication of Stewart et al. (2007), using a different methodology and analysis, and a larger sample size. We test the hypothesis that an ambiguity penalty occurs when the reader is motivated to process sentences at a deeper level, while an ambiguity advantage occurs under more shallow processing conditions as ambiguous pronouns may remain underspecified. If this is the case, a model of ambiguity resolution would be needed that accounts for both ambiguity advantages and penalties.

Ambiguity penalty

While ambiguous attachment sentences such as the ones used in Swets et al. (2008) are often complex and fairly unnatural, ambiguous pronominal references are found in everyday language and in less complex sentence constructions. A study by Stewart et al. (2007) examined the processing of ambiguous pronouns using a self-paced reading task. Their experimental design manipulated the ambiguity of the pronominal reference and the processing depth. Participants were presented with sentences that contained two referents and a personal pronoun that was either ambiguous, as in (1), or could refer to only one of the referents due to the presence of disambiguating gender information, as in (2) and (3).

(1) Paul$_i$ lent Rick$_j$ the CD before **he**$_{i/j}$ left for the holidays. (*ambiguous*)

(2) Paul$_i$ lent Kate$_j$ the CD before **he**$_i$ left for the holidays. (*first character reference*)

(3) Paul$_i$ lent Kate$_j$ the CD before **she**$_j$ left for the holidays. (*second character reference*)

Depth of processing was manipulated through offline comprehension questions. To probe *shallow* processing, participants were asked questions requiring a *yes/no* answer on only a third of the target sentences (e.g., *Did Paul lend Kate the CD*?) and a third of the filler items. Half of these questions targeted the information content of the main clause and half targeted the information content of the embedded clause. In the *deep* processing condition, questions on *all* experimental items and fillers targeted the information content of the embedded clause, and, hence, required resolution of the pronoun (e.g., *Who left for the holidays?*).

The results showed that in the shallow processing condition the total sentence reading times in the different conditions did not differ from each other (although numerically the ambiguous condition was read fastest). In the deep processing condition, ambiguous sentences were read more slowly than first and second character reference sentences (i.e., an ambiguity penalty or disadvantage). Stewart et al. (2007) proposed that readers invest processing resources to interpret the ambiguous pronoun when engaged in deep processing only, and not under shallow processing.

A second experiment, in which participants read sentences in a cumulative word-by-word self-paced reading paradigm paired only with shallow processing comprehension questions, again showed no effect of ambiguity on response times to the pronoun region and to the word after the pronoun in the first sentence. However, the critical sentences in this experiment were followed by a second sentence in which the ambiguous pronoun from the first sentence was disambiguated (e.g., for (1): *He went to the Bahamas and sent **Rick** a postcard from the hotel*, for which disambiguation occurs at the repeated name of the character to whom the pronoun does not refer). In the second sentence, a significant effect of ambiguity was found for the disambiguating region. The region that contained the repeated name was read more slowly in the ambiguous conditions than in the unambiguous conditions.

Based on the results of both experiments, Stewart et al. concluded that, under deep processing, readers attempt to assign a referent to the ambiguous pronoun even in a context where no disambiguation is available to indicate which assignment is correct. In contrast, under shallow processing, ambiguous pronouns would ultimately be interpreted but only when the reader is able to do this with certainty. The authors further argued that readers delayed the resolution of the ambiguous pronoun until disambiguating information was available, such that initial underspecification becomes more specified when disambiguating information is encountered, even under shallow processing.

Ambiguity advantage

In contrast to the results by Stewart et al. (2007), Grant et al. (2020) reported an ambiguity *advantage* for both structural and referential ambiguities in two experiments measuring eye movements during reading. We focus here on the results of the referential ambiguity study, for which the conditions are shown in (4) and (6).[1]

(4) We met the brother$_i$ of the waiter$_j$ when **he**$_{i/j}$ visited the restaurant, but we didn't talk for long. (*ambiguous*)

(5) We met the brother$_i$ of the waitress$_j$ when **he**$_i$ visited the restaurant, but we didn't talk for long. (*first character reference*)

(6) We met the sister$_j$ of the waiter$_i$ when **he**$_i$ visited the restaurant, but we didn't talk for long. (*second character reference*)

In the ambiguous condition, as in (4), both nouns (e.g., *brother* and *waiter*) were possible referents, while in the unambiguous conditions, as in (5) and (6), only one of the nouns was a possible referent. Rather than using personal names as in Stewart et al. (2007), gender information was based on a mix of lexical-semantic cues (e.g., *brother*), morphological cues (e.g., *waitress*), and stereotypical bias (e.g., *beautician*). The results showed an overall ambiguity advantage: similar to the results for structural ambiguities, Grant et al. reported significantly shorter reading times for ambiguous pronouns compared to unambiguous pronouns in go-past time (defined as the sum of all fixations from first entering a region until leaving it to the right).

[1] In Grant et al. (2020), the first character reference condition is referred to as 'high reference' and the second character reference condition is referred to as 'low reference'; we adopt the terminology used in Stewart et al. (2007) for consistency.

To rule out that the ambiguity advantage was due to the lack of salience for the referents, a second experiment was run, in which, similar to the design in Stewart et al. (2007), the main clause subject and object formed the potential referents. The referents were either definite descriptions that had definitional gender (e.g., *king*), or were proper names that were gender unambiguous (e.g., *Olivia*). The conditions were otherwise similar to those in the first experiment, as illustrated in (7) to (9).

(7) The young prince$_i$ showed the revered king$_j$ that **he**$_{i/j}$ would be a fine leader of the Tharassian empire. (*ambiguous*)

(8) The young prince$_i$ showed the revered queen$_j$ that **he**$_i$ would be a fine leader of the Tharassian empire. (*first character reference*)

(9) The revered queen$_j$ showed the young prince$_i$ that **he**$_i$ would be a fine leader of the Tharassian empire. (*second character reference*)

The results again revealed an ambiguity advantage effect in three out of five reading measures (first pass, go-past, and total time) in the post-critical region, with lower reading times for the ambiguous condition compared to either unambiguous condition. Grant et al. (2020) interpreted these results along the lines of an Unrestricted Race Model (cf. Van Gompel, Pickering, & Traxler, 2000) in which different referents ‘race’ to be selected as an antecedent for a pronoun, with the one that becomes available most quickly being adopted as the referent. If the antecedent that wins turns out to not match the pronoun’s gender features, the reference of the pronoun needs to be re-evaluated. The slower reading times in the unambiguous conditions reflect a penalty due to reanalysis: on some portion of trials, the processor selects the ultimately incorrect reference, which must then be reanalyzed. Such reanalysis never occurs for ambiguous trials as the parser will never be revealed to be wrong. Hence, reading times are predicted to be faster for ambiguous sentences compared to unambiguous ones.

In sum, Stewart et al. (2007) showed an ambiguity penalty when all sentences were followed by a targeted comprehension question in the deep processing condition (Experiment 1), and no significant difference between ambiguous and unambiguous sentences when 4/24 experimental trials were followed by a targeted comprehension question in the shallow processing condition (Experiment 1 and 2). Grant et al. (2020) showed ambiguity advantages when 12/30 sentences (Experiment 1) and 12/24 sentences (Experiment 2) were followed by a targeted comprehension question. However, these studies cannot be compared directly as different methodologies were used: Stewart et al. (2007, Experiment 1) analyzed whole-sentence self-paced reading times and Grant et al. (2020) analyzed eye-tracking measures while reading.

Depth of processing

There is some evidence suggesting that a less demanding task may lead comprehenders to underspecify certain syntactic relations. Ferreira and Yang (2019), for example, discuss how readers may flexibly adjust their reading strategies depending on their goals, and that the ultimate comprehension level may depend on reader or listener engagement or motivation. The results by Stewart et al. (2007) are suggestive in this respect, showing an ambiguity penalty only under deep processing conditions, and not under shallow processing. Swets et al. (2008), focusing on structural ambiguities, further showed that the processing of structurally ambiguous sentences is sensitive to depth of processing. They examined structural ambiguities in the form of sentences with a relative clause that could attach to one of two sites, as in (10), and sentences in which the relative clause occurred with a disambiguating reflexive pronoun, as in (11) and (12).

(10) The maid of the princess **who scratched herself** in public was terribly humiliated. (*ambiguous*)

(11) The son of the princess **who scratched himself** in public was terribly humiliated. (*high attachment*)

(12) The son of the princess **who scratched herself** in public was terribly humiliated. (*low attachment*)

Depth of processing was manipulated through the number and difficulty of the comprehension questions that followed the sentences. The results showed an ambiguity advantage only when comprehension questions were sparse and superficial, but an advantage for disambiguated low attachment sentences when comprehension questions were frequent and targeted the ambiguity.

Swets et al. argued that the comprehension system creates an underspecified representation when there is no basis on which to fully commit to an attachment decision in an ambiguous sentence. In other words, the syntactic representation in the globally ambiguous case may remain underspecified under shallow processing, but not under deep processing (see also Ferreira, Bailey, & Ferraro, 2002). Building on the results by Swets et al. (2008), Logačev and Vasishth (2016a) provided further evidence that, given a sufficiently high degree of task difficulty, sentences with attachment ambiguities are read more slowly than unambiguous sentences.

The current study

The goal of the present study was to examine to what extent the processing of referential ambiguities is modulated by processing depth, as manipulated through comprehension questions. If deep processing conditions result in an ambiguity penalty, and shallow processing conditions in an ambiguity advantage, this could potentially explain the discrepancy in earlier results on ambiguous referential processing (Grant et al., 2020; Stewart et al., 2007). To examine the effects of processing depth on the processing of referential ambiguities, we ran a conceptual replication of the first experiment in Stewart et al. (2007).

The materials consisted of Dutch sentences similar to the English sentences used in Stewart et al. (2007) and in Grant et al. (2020, Experiment 2). The sentences were ambiguous or disambiguated through gender features. In the ambiguous sentences, as in (13), the pronominal reference could refer to either character. In the disambiguated sentences, the pronoun could only refer to the gender-consistent referent, which was the first-mentioned character in the first character reference (14), or the second-mentioned in the second character reference (15). Sample items, including translations, are given below (a full stimulus list is provided as supplemental material; see Data Accessibility Statement).

(13) *Ambiguous*:
Abel$_i$ leende Gijs$_j$ het boek voor **hij**$_{i/j}$ op vakantie ging.
(Translation: Abel$_i$ lent Gijs$_j$ the book before he$_{i/j}$ went on holiday.)

(14) *First character reference*:
Abel$_i$ leende Zoë$_j$ het boek voor **hij**$_i$ op vakantie ging.
(Translation: Abel$_i$ lent Zoë$_j$ the book before he$_i$ went on holiday.)

(15) *Second character reference*:
Zoë$_i$ leende Abel$_j$ het boek voor **hij**$_j$ op vakantie ging.
(Translation: Zoë$_i$ lent Abel$_j$ the book before he$_j$ went on holiday.)

Processing depth was manipulated through the type of comprehension questions that participants were asked to answer. Sentences in the shallow processing condition were followed by questions that probed the information content of the main clause (e.g., *Wie leende het boek aan Gijs?* ‘Who lent the book to Gijs?’) and which did not require the participant to resolve the pronominal reference in the embedded clause. Sentences in the deep processing condition were followed by a question that probed the information content of the embedded clause and that could not be answered without resolving the pronominal reference (e.g., *Wie ging op vakantie?* ‘Who went on holiday?’). Note that we manipulated the required depth of processing of the embedded clause, not of the entire sentence. We return to this point in the Discussion.

If depth of processing modulates ambiguity resolution, we predict that participants show an ambiguity advantage under shallow processing conditions, but an ambiguity penalty under deep processing conditions. Hence, under deep processing, we predict that reading times for ambiguous pronouns would be longer than those for unambiguous pronouns as readers will attempt to resolve the reference. We predict shorter reading times for ambiguous pronouns under shallow processing as ambiguous pronominal references may remain underspecified under certain conditions, as was argued for English by Stewart et al. (2007).

We also examined question responses and question answering times to shed further light on the way in which pronominal resolution occurs, testing whether readers used a heuristic that favored one referent over the other. Referents in subject position are often argued to hold a privileged status, which predicts that the first-mentioned character may be more often co-indexed with the pronoun, even in the absence of disambiguating information (e.g. Arnold, 1998; Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; McDonald & MacWhinney, 1995; Stewart et al., 2007; Van Rij, Van Rijn, & Hendriks, 2013).

Our study improved on the original design in Stewart et al. (2007, Experiment 1) in several ways, providing a different methodology, a larger sample size, and up-to-date statistical analyses. We used a *word-by-word* self-paced reading task (Just, Carpenter, & Woolley, 1982), which allowed us to localize the effects and obtain a more detailed picture of the time course of processing, compared to an analysis based on whole-sentence reading times in Stewart et al. (2007). In addition, we used a non-cumulative (moving window) paradigm, instead of a cumulative reading task. The cumulative paradigm has been shown to be problematic for studying on-line parsing, as participants develop a reading strategy in which they reveal several segments of a stimulus at a time and then read them all at once, or even wait to process a sentence until it is entirely revealed (Ferreira & Henderson, 1990; Just et al., 1982).

Finally, given the fact that the replication crisis has surfaced also in the pronoun interpretation literature (see e.g., Chow, Lewis, and Phillips 2014's failed attempts to replicate the findings in Badecker and Straub 2002), we included more experimental items. While the original study included 24 target sentences, we included a total of 60 target sentences. We performed an a-priori power analysis to determine the sample needed to observe an effect with at least 80% power (see below). While it is possible that the lack of a significant ambiguity advantage in the shallow processing condition in Stewart et al. (2007, Experiment 1) was due to a lack of power, our experiment should have appropriate power to detect the effect. We further used linear mixed-effects models to analyze the data, rather than traditional ANOVAs.

Methods

Participants

Participants were 156 non-dyslexic native speakers of Dutch who reported having normal or corrected-to-normal vision (mean age = 24.9; sd = 4.36). Participants were recruited from the participant pool at the Max Planck Institute for Psycholinguistics, and anyone who took part in the norming study was excluded from participation. Participants received a payment of €6 for their participation. Informed consent was obtained from each participant. The study was approved by the ethics board of the Faculty of Social Sciences of Radboud University.

The number of participants was determined through an a-priori power analysis conducted in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), which suggested that a sample size of 39 participants (13 per list) would give us at least 80% power to observe a main effect of sentence type with an effect size *f* of 0.15 or larger (which was calculated based on the results in Stewart et al. 2007, Experiment 2).[2] To account for the interaction with processing depth, we included twice as many participants per cell, hence, including four times as many participants, leading to a total of 156 participants.

Materials and design

As in Stewart et al. (2007), we used a 2 × 3 mixed design, consisting of a between-participants factor 'depth of processing' of the embedded clause with two levels (shallow processing, deep processing), and a within-participants factor 'sentence type' with three levels (ambiguous, first character reference, second character reference). The materials consisted of 60 target sentences, with an ambiguous version, a first character reference version, and a second character reference version each. Because self-paced reading experiments typically show spill-over effects, in that difficulty induced by a given word slows down the reading times of subsequent words, the sentences were created such that the word immediately following the pronoun (PRONOUN + 1) was always a closed class item (a preposition or determiner) and the word after that (PRONOUN + 2) was an open class item. We made sure that the frequency of the open class items was not extremely low (i.e., >1 per million words) or extremely high (≤ 100 per million). The mean word frequency of the PRONOUN + 1 items was 17862.88 per million words (sd = 8345.13); the mean frequency of PRONOUN + 2 items was 33.65 per million (sd = 29.28). Word frequencies were retrieved from the SubtLex-NL database (Keuleers, Brysbaert, & New, 2010).

We ensured that the critical regions of interest in the sentences in the different conditions were identical. To do so, we switched the order of the female and male proper names in the first and second character reference sentences (e.g., *Abel lent Zoë* and *Zoë lent Abel*) which ensured the use of the same pronoun across conditions. All of the critical sentences occurred with the third person masculine personal pronoun *hij* 'he', as the Dutch feminine pronoun *zij* or *ze* 'she' can also refer to the third person plural pronoun 'they', in which case a sentence as in (13) would be disambiguated only at the sentence-final verb. Dutch stereotypically male or female proper names were selected from the Dutch First Name database (Nederlandse Voornamenbank[3]). The different versions (ambiguous, first character reference, second character reference) of each experimental item were rotated across three lists, so that each participant saw every item only once. We further included 30 filler sentences that also consisted of a main and an embedded clause, but differed from the experimental items in that they only contained one character.

[2] G*Power uses Cohen's *f*, which is an effect size appropriate for a within-subjects repeated measures analysis of variance (ANOVA). We used this effect size in our power calculation despite using mixed-effect models for analysis as the two types of analyses are conceptually similar.

[3] https://www.meertens.knaw.nl/nvb/.

Twenty of the filler sentences occurred with the feminine personal pronoun *ze* ‘she’ and ten occurred without any personal pronoun or with the masculine personal pronoun *hij* ‘he’. Filler items were identical for each of the lists.

To ensure that participants could not use plausibility information to guide the interpretation of the pronoun in the critical sentences, we conducted a norming study to establish that the first and second character reference interpretations were equally plausible. The norming materials consisted of 70 unambiguous sentences. Following the procedure reported in Stewart et al. (2007), a first character reference and a second character reference version were included for each sentence. The norming study consisted of two lists, with first and second character references of the same sentence occurring on different lists. Each list contained equal numbers of first and second character reference sentences. In addition, we included 20 sentences that were implausible on the basis of pragmatic or world knowledge (e.g., *Paul lachte Sophie uit toen hij van zijn fiets viel* ‘Paul laughed at Sophie when he fell off his bike’ and *Rick stuurde een onderzeeër om de bosbrand te blussen* ‘Rick sent a submarine to extinguish the forest fire’). The norming study was run online, using an online tool for web experiments (Frinex) developed by the technical group of the MPI.

A total of 40 native speakers of Dutch ($n = 20$ per list; mean age = 23.28; sd = 2.92) were asked to rate each sentence for plausibility of the event it described (*Hoe waarschijnlijk is het dat dit gebeurt?* ‘How likely is this happening?’) on a 7-point Likert scale, with 1 corresponding to ‘highly implausible’ (*zeer onwaarschijnlijk* in Dutch) and 7 to ‘highly plausible’ (*zeer waarschijnlijk*). Based on the results of the norming study, a set of 60 sentences were selected for which the difference between the mean ratings for the two versions of the same sentence was minimal. The mean difference between the first and second character reference sentences for these 60 sentences was 0.40 (sd = 0.36), with a range of 0 to 1.2. The mean plausibility of the sentences referring to the first-mentioned character was 5.45 (sd = 0.84), and the mean plausibility of the sentences referring to the second-mentioned character was 5.39 (sd = 0.83).

All sentences in the main experiment were followed by a two-choice comprehension question. As described above, these questions served to manipulated the processing depth. In the shallow processing condition, all questions probed the information content of the main clause. Hence, the embedded clause was irrelevant to the comprehension question and participants were not required to resolve the pronominal reference. Sentence (13) above, for instance, was followed by the question *Wie leende het boek aan Gijs?* ‘Who lent the book to Gijs?’, with answer options *Abel* (correct) or *Zoë* (incorrect). To answer this question, participants were merely required to remember who the characters were in the sentence, since only one of the two answer options occurred in the sentence. A third of the sentences were followed by a question that did not mention any of the characters (e.g., *Wie werd er uitgelachen?* ‘Who was laughed at?’ after the sentence *Jordy lachte Marc uit toen hij de grap begreep* ‘Jordy laughed at Marc when he understood the joke'), with answer options *Jordy* (incorrect) and *Marc* (correct). In these cases, participants needed to remember who did what as expressed in the main clause, but they, crucially, did not need to resolve the pronominal reference in the embedded clause. Half of the shallow processing questions targeted the first-mentioned referent, and half targeted the second-mentioned referent (i.e., *Aan wie leende Abel het boek?* 'To whom did Abel lend the book?').

In the deep processing condition, all sentences were followed by a question that probed the information content of the embedded clause, such that the questions could not be answered without resolving the pronominal reference. Sentence (13) above was paired with the question *Wie ging op vakantie?* 'Who went on holiday?' with answer choices *Abel* and *Gijs*. In the ambiguous condition, both answers reflected potential interpretations of the pronoun, while in the disambiguated sentences only one of the two choices was correct (e.g., for (18): *Abel* (correct) or *Zoë* (incorrect)).

Procedure

The experiment was run online, using the same online tool for web experiments as was used for the norming study (Frinex). In the experiment, participants read sentences in a word-by-word self-paced reading paradigm (Just et al., 1982). Participants used their own keyboards to respond. Upon pressing the space bar, a trial began and the first word of the sentence appeared. Participants continued to press the space bar to read each successive word, and as each word appeared, preceding words were masked again. Participants were instructed to read at a natural pace and to make sure they understood what they were reading so that they could answer the comprehension questions accurately. Reading times were measured for each word from the time it appeared on the screen until the space bar was pressed.

After participants read the complete sentence, a comprehension question appeared on the screen with two answer choices, to which participants responded with a key press. Participants were randomly assigned to either the shallow processing or the deep processing condition. In the shallow processing condition, the answers consisted of the correct character name and a character name that had not appeared in the preceding sentence. In a portion of the questions no character was mentioned in the question (e.g., *Wie vertelde over het probleem?* 'Who mentioned the problem?'), in which case the two character names from the sentence were given as possible answers. In the deep processing condition, the two character names that had appeared in the preceding sentence were always depicted on the screen as possible answers. The positions of the two answers on the screen were randomized. After answering, participants proceeded to the next trial by pressing the space bar.

We generated twelve pseudo-random orders of item presentation to which participants were randomly assigned. The experimental session was preceded by a set of 8 practice trials. The practice trials were meant to familiarize the participant with the experiment and to mimic the different conditions of the between-participants manipulation of processing depth: in the shallow processing condition, practice trials were followed by questions targeting main clause information, while questions targeted the information content of the embedded clause were used in the deep processing condition.

Analysis

We recorded reading times (RTs) for each word, and accuracy and question-answering times for the comprehension questions after each sentence. Eight participants were removed for technical reasons (only part of their responses were properly recorded, likely due to connection issues or because they continuously pressed the space bar for at least a portion of the experiment). We then calculated the mean accuracy on the comprehension questions after filler sentences for each participant, and removed one participant with $<75\%$ accuracy from further analyses. All remaining 147 participants had a high accuracy on the filler items (on average 99% accurate on the filler items under both deep and shallow processing conditions).

The reading and question-answering time data were analyzed with linear mixed-effects models, using the `lme4` package (Bates, Kliegl, Vasishth, & Baayen, 2015) in the `R` environment (R Core Team, 2016) (version 3.6.0). Post-hoc contrasts were performed using the `emmeans` package (Lenth, Singmann, Love, Buerkner, & Herve, 2018) and *p*-values were computed using the Satterthwaite approximation for degrees of freedom, as implemented in the `lmerTest` package (Kuznetsova, Brockhoff, & Christensen, 2016). We used the optimiser "bobyqa" to help with convergence issues.

Analysis of reading times

Unreasonably short and long latencies (<100 ms and >2500 ms) were excluded prior to analysis, which resulted in a loss of less than 1% of the data. No additional outlier removal was performed. Reading times were log-transformed (natural log) to normalize residuals and reduce the effect of extreme data points. Log-transformed reading times were then regressed against two factors that are known to affect reading times in self-paced reading tasks: word length and the list position of the sentence in the stimulus list (i.e., longer words predict longer reading times and later list positions predict faster reading times; Hofmeister, 2011; Hofmeister & Vasishth, 2014). The model estimating these effects included by-participant slopes of WORD LENGTH and LIST POSITION. Both WORD LENGTH and LIST POSITION were log-transformed with a natural logarithm and then centered. The data from all sentences (experimental and filler) were used to produce maximally general estimates (Hofmeister & Vasishth, 2014).

We analyzed the residual log reading times in five separate regions of interest: the pronoun itself (PRONOUN), the two regions immediately following the pronoun (PRONOUN + 1 and PRONOUN + 2), and the two regions preceding it (PRONOUN-1 and PRONOUN-2). The words used in each of these regions were identical across conditions. We included the two words following the pronoun as self-paced reading experiments often show spill-over effects, i.e., effects on words following the region of interest in the sentence. To establish whether there were any differences between the conditions prior to the pronoun, we further included the two words preceding the pronoun: the PRONOUN-1 region (which was always the first word of the embedded clause) and the PRONOUN-2 region. Linear mixed-effects models for the different regions of interest were fitted with fixed effects of SENTENCE TYPE and DEPTH OF PROCESSING and their interactions. SENTENCE TYPE was coded with successive differences contrast coding (using the `contr.sdif()` function in the MASS package; Venables & Ripley, 2002), comparing the differences between the means of the second and first levels (ambiguous vs. first character reference), and the third and second levels (second character reference vs. ambiguous). DEPTH OF PROCESSING was coded with scaled sum contrasts (or effect coding; –0.5 and 0.5). We fitted the models with the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013), which included by-participant and by-item intercepts, by-participant slopes for SENTENCE TYPE, and by-item slopes for SENTENCE TYPE, DEPTH, and their interactions. Random slopes were removed when the model did not converge; random slopes correlated above 0.95 were also removed to avoid overfitting. This resulted in models that included by-participant and by-item intercepts and a by-item slope for DEPTH in the PRONOUN, PRONOUN + 1, and PRONOUN + 2 regions, and models with by-participant and by-item intercepts and by-item slopes for SENTENCE TYPE, DEPTH, and their interactions for the PRONOUN-1 and PRONOUN-2 regions.

Analysis of question responses

For the question-answering times, unreasonably high (> 10000 ms) and low (< 700 ms) response times were removed. This resulted in 0.85% of the response time data being excluded. Response times were log-transformed (natural log) and log response times more than 2.5 standard deviations from the condition mean (per participant) were excluded, resulting in the removal of a further 1.22% of the data. Log-transformed question-answering times were analyzed with linear mixed-effects models that included a fixed effect of SENTENCE TYPE and DEPTH OF PROCESSING and their interactions. As with the analysis of reading times, SENTENCE TYPE was coded with successive differences contrast coding, comparing the ambiguous condition to the first character reference and to the second character reference conditions. DEPTH OF PROCESSING was coded with scaled sum contrasts and random intercepts for subjects and items were included.

Results

Reading times

A summary of the non-transformed reading times across the experimental conditions is given in **Table 1**. Analyses were conducted on residual log reading times, as shown in **Figure 1**. **Table 2** gives the output of the mixed-effect models.

Depth	Sentence type	PRONOUN-2	PRONOUN-1	PRONOUN	PRONOUN + 1	PRONOUN + 2
Shallow	Ambiguous	309 (4.82)	313 (4.95)	280 (3.52)	273 (3.21)	285 (3.50)
	First	314 (5.10)	319 (4.81)	285 (3.29)	272 (3.05)	288 (3.31)
	Second	309 (5.00)	310 (4.86)	283 (3.31)	276 (3.23)	285 (3.71)
Deep	Ambiguous	364 (5.76)	377 (5.57)	350 (4.51)	335 (4.18)	355 (4.41)
	First	368 (5.69)	366 (5.16)	345 (4.34)	326 (3.50)	338 (3.77)
	Second	367 (5.83)	365 (5.00)	351 (5.05)	330 (3.69)	345 (4.23)

Table 1: Means (and SEs) for reading latencies in the three regions of interest under shallow and deep processing conditions.

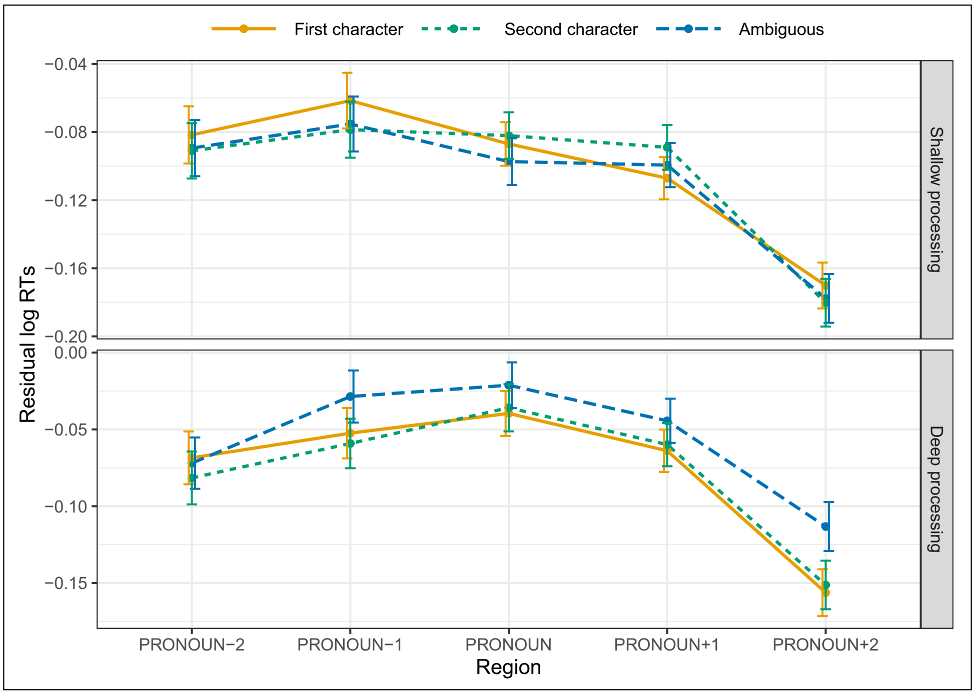

Figure 1: Residual log reading times in the regions of interest under shallow and deep processing conditions (regressed against word length and the log list position). Error bars represent 95% confidence intervals.

	PRONOUN-2		PRONOUN-1		PRONOUN		PRONOUN + 1		PRONOUN + 2	
	***β* (SE)**	***p*-value**	***β* (SE)**	***p*-value**	***β* (SE)**	***p*-value**	***β* (SE)**	***p*-value**	***β* (SE)**	***p*-value**
Intercept	–0.081 (0.012)	< **0.001**	–0.060 (0.011)	< **0.001**	–0.060 (0.008)	< **0.001**	–0.077 (0.008)	< **0.001**	–0.158 (0.009)	< **0.001**
First vs. Amb	–0.006 (0.010)	0.577	0.005 (0.009)	0.540	0.005 (0.007)	0.478	0.014 (0.007)	**0.033**	0.019 (0.007)	**0.008**
Amb vs. Sec	–0.005 (0.009)	0.596	–0.015 (0.010)	0.127	0.001 (0.007)	0.893	–0.002 (0.007)	0.763	–0.020 (0.007)	**0.006**
Depth	–0.014 (0.013)	0.280	–0.026 (0.014)	0.069	–0.057 (0.012)	< **0.001**	–0.043 (0.011)	< **0.001**	–0.037 (0.014)	**0.012**
First vs. Amb × Depth	–0.004 (0.017)	0.804	–0.038 (0.019)	0.053	–0.030 (0.014)	**0.033**	–0.011 (0.013)	0.393	–0.050 (0.014)	< **0.001**
Amb vs. Sec × Depth	0.007 (0.017)	0.667	0.029 (0.020)	0.159	0.031 (0.014)	**0.026**	0.026 (0.013)	0.053	0.035 (0.014)	**0.013**

Table 2: Fixed effect estimates, standard errors, and p-values for linear mixed effects regression for the residual log reading times. Results from the full model are given there was a marked change in fixed effects in the reduced model.

Throughout the PRONOUN region and the two regions following it, reading times were strongly affected by processing depth. The shallow processing condition elicited faster average reading times (280 ms on average) than the deep processing condition (342 ms on average) when comparing reading times in the PRONOUN region and the two spill-over regions. A main effect of DEPTH was found for each of these three regions (PRONOUN region: $\beta = –0.057$, $p < 0.001$; PRONOUN + 1 region: $\beta = –0.043$, $p < 0.001$; PRONOUN + 2 region: $\beta= –0.037$, $p = 0.012$). When considering the whole-sentence reading times, the deep processing condition also led to longer average RTs (4746 ms) compared to the shallow processing condition (3950 ms). However, the effect of DEPTH was only marginally significant in the region immediately preceding the pronoun (PRONOUN-1 region: $\beta = –0.026$, $p = 0.069$) and was not significant in the region before that (PRONOUN-2 region: $p = 0.280$). This indicates that participants who were asked 'superficial' questions in our shallow processing condition read the embedded clause faster than participants who were asked questions that specifically targeted the information content of that embedded clause. While the manipulation of processing depth affected the processing of the embedded clause (which started at the word in the PRONOUN-1 region), this further shows that it did not necessarily affect the generic processing depth over the entire sentence.

Zooming in on the reading times per region of interest, the results in the PRONOUN region did not reveal a main effect of SENTENCE TYPE when comparing the ambiguous condition to the first character ($p = 0.478$) and second character reference ($p = 0.893$) conditions. Crucial for the aim of the current experiment are the interactions between SENTENCE TYPE and DEPTH. For the PRONOUN region, these interactions were significant when comparing ambiguous and first character reference sentences ($\beta = –0.030$, $p = 0.033$) and when comparing second character reference and ambiguous sentences ($\beta = 0.031$, $p = 0.026$).[4] This suggests that there is an effect of processing depth (or task demand) on the difference in reading times between ambiguous and unambiguous sentences in this region, with shorter reading times for ambiguous sentences under shallow processing but longer reading times under deep processing, which would be in line with earlier findings of an ambiguity advantage. Numerically, the ambiguous condition was read faster under shallow processing compared to the unambiguous conditions. However, post-hoc comparisons (with a correction for multiple comparisons using the Tukey method), revealed no significant differences in reading times between the different sentence types under shallow processing (first-second: $p = 0.787$; first-ambiguous: $p = 0.575$; ambiguous-second: $p = 0.219$) or under deep processing (first-second: $p = 0.855$; first-ambiguous: $p = 0.111$; ambiguous-second: $p = 0.304$).

The results in the PRONOUN + 1 region, which consisted of highly frequent words such as prepositions, showed a difference between the reading times in the ambiguous and first character

[4] In the full model, the interactions between SENTENCE TYPE and DEPTH in the PRONOUN region were only marginally significant (ambiguous vs. first character: $\beta = –0.029$, $p = 0.069$; second character vs. ambiguous: $\beta = 0.031$, $p = 0.062$).

reference conditions across levels of DEPTH ($\beta = 0.014$, $p = 0.033$).[5] This suggests that the reading times for the first character reference sentences were faster compared to the ambiguous sentences when looking at both processing depths. However, and as shown in **Table 2**, none of the interactions between SENTENCE TYPE and DEPTH were significant.

The PRONOUN + 2 region showed a main effect of SENTENCE TYPE when comparing reading times for ambiguous to the first character reference ($\beta = 0.019$, $p = 0.008$) and the second character reference sentences ($\beta = –0.020$, $p = 0.006$). This shows that, across levels of DEPTH, ambiguous sentences were read more slowly than either of the unambiguous ones, although this effect is probably driven by the deep processing condition as discussed below. The interaction between SENTENCE TYPE and DEPTH was significant when comparing the ambiguous and the first character reference sentences ($\beta = –0.050$, $p < 0.001$) and the ambiguous and second character reference sentences ($\beta = 0.035$, $p = 0.013$).[6] This again suggests that the difference in reading times between ambiguous and unambiguous sentences depends on the depth of processing that participants engaged in. Post-hoc comparisons indeed showed that the ambiguous sentences were read significantly more slowly than the unambiguous sentences under deep processing (first-ambiguous: $\beta = –0.044$, $p < 0.001$; ambiguous-second: $\beta = 0.037$, $p < 0.001$). No differences were found between the unambiguous conditions ($p = 0.794$). The longer reading times for ambiguous sentences reflect an ambiguity disadvantage or penalty under deep processing. Under shallow processing, no significant differences between any of the sentence types were found (first-ambiguous: $p = 0.817$; first-second: $p = 0.704$; ambiguous-second: $p = 0.980$).

Finally, we analyzed the two regions preceding the pronoun which, similar to the regions following the pronoun, were identical across conditions. In the PRONOUN-1 region, the model revealed a marginally significant interaction when comparing the ambiguous sentences to the first character reference sentences ($\beta = –0.038$, $p = 0.053$), in the same direction as the interaction in the PRONOUN region. Although this interaction was only marginally significant (in both the full and reduced models) and should be interpreted with caution, this could indicate that participants may have learned over the course of the experiment that an ambiguous pronoun may follow after having seen two stereotypically male proper names in the main clause (as opposed to one male and one female name). No other effects were significant. The model for the PRONOUN-2 region revealed no significant effects (see **Table 2**).

Question responses

The analysis of the question responses showed high accuracy rates across conditions, as summarized in **Table 3**. When examining answers to questions after ambiguous sentences in the

[5] This difference was only marginally significant in the full model ($p = 0.058$).

[6] The interaction when comparing the ambiguous and second character reference sentences was only marginally significant in the full model ($p = 0.083$).

deep processing condition (for which participants had to resolve the pronominal reference), a strong preference to resolve the ambiguous pronominal reference to the first character reference was observed: out of 1460 answers, 1040 (71%) referred to the first character and 420 (29%) to the second character (**Figure 2A**). This is in line with the consistent finding of a tendency for the referent of the subject or first-mentioned noun phrase to be more accessible than other entities (e.g. Arnold, 1998; Arnold et al., 2000; McDonald & MacWhinney, 1995; Van Rij et al., 2013). It is likely that the fact that referents in subject position are more prominent or salient resulted in this subject bias when resolving pronoun reference.

Depth	**Sentence type**	**Accuracy**	
Shallow	First	0.95	(0.006)
	Second	0.94	(0.006)
	Ambiguous	0.94	(0.006)
	Filler	0.99	(0.002)
Deep	First	0.97	(0.004)
	Second	0.97	(0.004)
	Filler	0.99	(0.002)

Table 3: Accuracy (and SEs) per sentence type and under shallow and deep processing conditions.

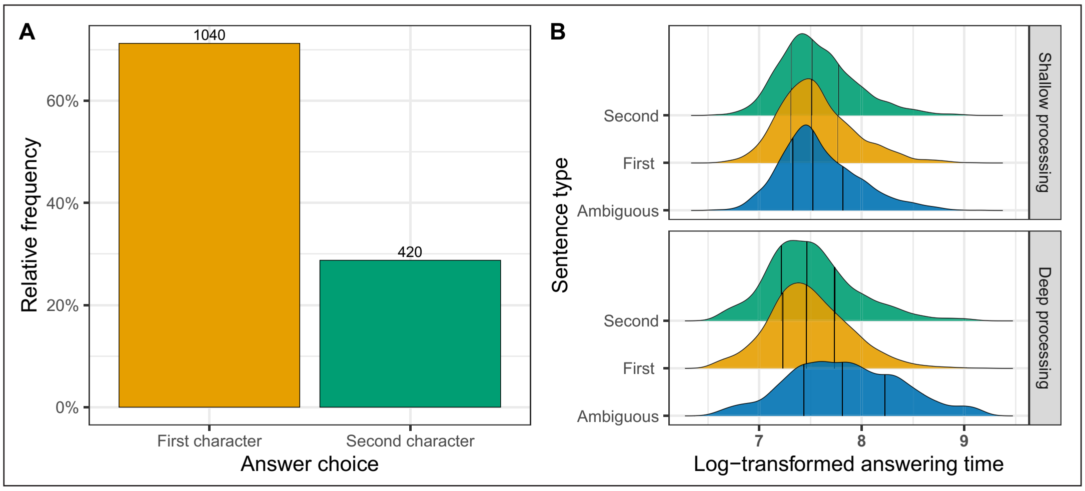

Figure 2: Question responses. **A:** Relative frequency of answers referring to either the first character or the second character given by participants after ambiguous sentences in the deep processing condition. **B:** Log-transformed (natural log) question answering times for the different sentence types under deep and shallow processing; the lines correspond to the first, second, and third quartile.

The analysis of question answering times (see **Figure 2B** and **Table 4** for the complete model output) showed a difference between the ambiguous and the first character reference conditions ($\beta = 0.184$; $p < 0.001$) and between the ambiguous and the second character reference conditions ($\beta = 0.172$; $p < 0.001$) across levels of DEPTH. DEPTH did not result in a significant main effect ($p = 0.292$). However, the interactions between SENTENCE TYPE and DEPTH were significant and showed a larger difference between ambiguous and first character reference sentences ($\beta = 0.318$; $p < 0.001$) and between ambiguous and second character reference sentences ($\beta = 0.308$; $p < 0.001$) under deep processing compared to shallow processing. Post-hoc comparisons (with a correction for multiple comparisons using the Tukey method) indeed showed no significant differences between sentence types under shallow processing (first-second: $p = 0.838$; first-ambiguous: $p = 0.124$; second-ambiguous: $p = 0.347$). Under deep processing conditions, however, participants took significantly longer to answer comprehension questions after ambiguous sentences compared to both first character reference ($\beta = 0.344$, $p < 0.001$) and second character reference sentences ($\beta = 0.326$, $p < 0.001$). No difference between the unambiguous sentences was found ($p = 0.341$).

	Question answering times			
	Estimate	**SE**	***t*-value**	***p*-value**
Intercept	7.593	0.023	336.119	<**0.001**
First vs. Amb	0.185	0.009	20.030	<**0.001**
Amb vs. Sec	–0.172	0.009	–18.653	<**0.001**
Depth	–0.037	0.035	–1.057	0.292
First vs. Amb × Depth	–0.319	0.018	–17.319	<**0.001**
Amb vs. Sec × Depth	0.308	0.018	16.736	<**0.001**

Table 4: Fixed effect estimates for linear mixed effects regression for log-transformed (natural log) question answering times.
Note. First = first character reference condition; Sec = second character reference condition; Amb = ambiguous condition.

General discussion

In previous studies, the processing of different types of ambiguities has been argued to result in either an *ambiguity advantage* or in an *ambiguity penalty*. The present study contributes to this line of research by examining why ambiguous sentences sometimes lead to ambiguity advantages and other times to ambiguity penalties. We tested the hypothesis that an ambiguity penalty occurs when task demands require a ‘deeper’ interpretation of the embedded clause and the pronoun in it, while an ambiguity advantage occurs when task demands allow for more superficial interpretations of the pronominal reference. To do so, we provided a conceptual replication of the experiments reported in Stewart et al. (2007).

The results showed a large effect of depth of processing of the embedded clause, with a reading time difference between the shallow and deep processing conditions of about 60 ms on average in the pronoun region and the two regions following it. Participants who were asked comprehension questions that were only superficial, in that they did not target the content of the embedded clause and did not require resolving the pronominal reference (the shallow processing condition), showed faster reading times in the regions of interest than participants who were asked questions that targeted the information content of the embedded clause and that could not be answered without resolving the pronominal reference (deep processing). This suggests that the questions in our deep processing condition did indeed lead to ‘deeper’ processing: participants read the relevant parts of the sentences more slowly, and hence likely more carefully, when they expected more difficult questions. The effect of depth was not significant in the PRONOUN-2 region (i.e., the region two words before the pronoun) and only marginally significant in the PRONOUN-1 region (i.e., the region immediately preceding the pronoun), which suggests that our manipulation of processing depth mainly affected the processing of the embedded clause, and not the processing depth of the entire sentence. This is most likely a consequence of the way in which we manipulated processing depth, as the comprehension questions manipulated the relevance of the embedded clause. In other words, participants read the embedded clauses, specifically on and following the critical pronouns, more slowly under deep processing, but not necessarily the entire sentences.

The most important results from the reading time analysis for the different regions were the significant interactions between sentence type and processing depth in the PRONOUN and PRONOUN + 2 regions. In these two regions, we found that the difference in reading times for ambiguous versus unambiguous sentences depended on the depth of processing.[7] Post-hoc comparisons revealed a significant *ambiguity penalty* (or *disadvantage*) in the PRONOUN + 2 region under deep processing, but not under shallow processing. In the PRONOUN region, the results for shallow processing showed numerically faster reading times for the ambiguous condition relative to the unambiguous conditions, which is in line with earlier findings of an *ambiguity advantage* under shallow conditions. However, this effect was not significant.

[7] We acknowledge that the marginal interaction in the pre-critical region points to the possibility that the ambiguity penalty for pronouns in the deep processing condition could, at least in part, be caused by participants learning that an ambiguous pronoun may follow two male proper names. As pointed out by a reviewer, if this were the case, the difference in processing behaviour between deep and shallow conditions may, at least in part, be an anticipatory effect. We leave this for further research.

Comparison to earlier findings

Our reading time results replicate the results reported in Stewart et al. (2007, Experiment 1), using a different methodology (i.e., measuring non-cumulative word-by-word reading times rather than whole-sentence reading times), a larger sample size to provide adequate power, up-to-date statistical analyses, and a language that is different from but related to English.

Stewart et al. (2007) found that whole-sentence self-paced reading times in the ambiguous and unambiguous conditions did not differ significantly from each other under shallow processing, while ambiguous sentences were read more slowly (i.e., an ambiguity penalty) than unambiguous sentences under deep processing. These results for whole-sentence reading times showed a similar pattern to our results for word-by-word reading times: in our results, the conditions also did not differ from each under under shallow processing, while an ambiguity penalty was found under deep processing. In contrast, Grant et al. (2020) observed a significant ambiguity advantage in the post-critical region when comparing sentences with an ambiguous pronominal reference to unambiguous sentences in an eye-tracking while reading experiment.

A relevant question is why we did not find a significant ambiguity advantage under shallow processing. One potential reason is that participants in our study, and in the experiments reported in Stewart et al. (2007), did not read the embedded clause well enough to lead to differences between the sentence types. Since, in our study, none of the comprehension questions in the shallow processing version targeted the content of the embedded clause, this may have lead some participants to only skim this part of the sentence. However, considering the fact that this was a self-paced reading experiment in which participants were presented each word of the sentence in a cumulative way (different from an eye-tracking experiment in which it is easier to skip over parts of the sentence), this seems unlikely. Moreover, Stewart et al. (2007) did include some (4/24) comprehension questions that targeted the embedded clause and did not find a significant ambiguity advantage either.

A different explanation relating to the shallow comprehension questions, suggested to us by a reviewer, is that our 'shallow' questions may not have been shallow enough to observe an ambiguity advantage. While none of the shallow questions targeted the pronominal reference, the questions did refer to the referents (asking who did something). This may have invited strategic attention to the referents, which may have wiped out a possible ambiguity advantage. In their study, Grant et al. (2020) also included questions that probed the comprehension of the pronoun, but did so after only a proportion of the trials. It could, therefore, be the case that the occurrence of a question relating to a referent on all trials motivated a less superficial type of referential processing even in the shallow condition, resulting in the absence of a significant ambiguity advantage.

It could also be the case that the processing system handles the resolution of structural (i.e., attachment) ambiguities and pronominal reference ambiguities differently. Unlike structural/

attachment ambiguities, which have typically been shown to facilitate processing (e.g. Clifton & Staub, 2008; Traxler et al., 1998; Van Gompel et al., 2005, 2001), referential ambiguities would then more often *hinder* processing (Badecker & Straub, 2002; Stewart et al., 2007). While this position is not unreasonable considering that referential and structural ambiguities are distinct phenomena, the findings by Grant et al. (2020) suggests that, when compared directly and under equivalent conditions, structural and referential ambiguity show similar patterns in eye movements during reading. This suggests that readers react to referential and structural ambiguity in a similar fashion.

Another possible explanation is that the ambiguity advantage for pronominal references may constitute such a small effect that even more statistical power and/or a more sensitive experimental paradigm is needed to observe it. For instance, it could be that the effect is more pronounced in an eye-tracking study, explaining why Grant et al. (2020) did find a significant ambiguity advantage.

Moving back to comparisons to earlier findings, we used a word-by-word self-paced reading task, while Stewart et al. (2007, Experiment 1) examined whole-sentence reading times. This allowed us to localize the effects and obtain a more detailed picture of the time course of processing. The results showed an ambiguity disadvantage in the PRONOUN + 2 region, which was a post-critical region. While spill-over effects are expected in self-paced reading experiments, this could potentially indicate that referential ambiguity does not immediately result in an ambiguity penalty, but that this effect is delayed. This is similar to what Logačev and Vasishth (2016a) reported for an ambiguity disadvantage in structurally ambiguous sentences, where a relatively late ambiguity disadvantage was observed at the last word of the relative clause. It is also in line with Stewart et al. (2007), who argued that readers may delay the resolution of ambiguous pronouns. A different line of literature on the processing of gender features also suggests that ambiguous pronouns do not disrupt comprehension immediately (e.g. Greene, McKoon, & Ratcliff, 1992; Rigalleau, Caplan, & Baudiffier, 2004), but may do so after a delay.

Finally, the question-answering results showed a larger difference between the ambiguous and both of the unambiguous conditions when questions required resolving the pronominal reference compared to when questions were only superficial. Similar results were reported in Stewart et al. (2007) and Swets et al. (2008), who showed that question-answering times were significantly longer after ambiguous than after unambiguous sentences when ‘deep’ processing questions were asked, but no such difference was found when questions were only superficial.

Theoretical implications

Our results provide additional evidence, along the lines of Swets et al. (2008) and Stewart et al. (2007), that the difference in reading times between referentially ambiguous and unambiguous sentences depends on the readers’ goals, in our case whether the manipulation of ‘depth of processing' required the reader to answer questions that targeted the antecedent of the pronoun in the embedded clause. This has important implications for models of sentence processing. In particular, it suggests that ambiguity resolution is a *flexible* process, such that the way in which ambiguous sentences are processed (relative to unambiguous ones) depends on the processing depth or task demands.

In previous studies of structural and referential ambiguities, ambiguity penalties have been taken as evidence for constraint-based parsing models or competition-based models, in which the parallel consideration of multiple structures gives rise to competition effects in processing (e.g., Badecker & Straub, 2002; MacDonald, 1994). In contrast, ambiguity advantages have been taken as support for non-competitive models such as the Unrestricted Race Model (e.g. Van Gompel et al., 2000), according to which potential syntactic analyses engage in a race before a single analysis is adopted. In a race-based model, disambiguated sentences are predicted to take longer to process than ambiguous ones because processing delays occur when subsequent information is inconsistent with a previously adopted analysis and the initial analysis needs to be revised (Van Gompel et al., 2000). Alternatively, an ambiguity advantage is predicted because the adopted analysis is the one that is completed the fastest, leading to shorter times on average when there are several candidates, as in globally ambiguous cases compared to unambiguous cases (Logačev & Vasishth, 2016a). The current results warrant the formulation of a model of ambiguity resolution that takes into account the effect of processing depth on ambiguity resolution, and that should be able to capture *both* ambiguity advantages under shallow processing and ambiguity penalties under deep processing.

One proposal that has been put forward to account for moderating effects of processing depth or task demands is based on *underspecification* (cf. Ferreira et al., 2002; Karimi & Ferreira, 2016; Stewart et al., 2007; Swets et al., 2008). This proposal is in line with the 'good-enough' processing approach, which argues that readers or listeners may construct sentence representations that are only 'good-enough', i.e., not necessarily complete and fully specified, but that suffice to accomplish the communicative goal at hand (e.g., Christianson, 2016; Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira & Lowder, 2016; Ferreira & Patson, 2007). For structural ambiguities, Swets et al. (2008) argued that readers may strategically underspecify the representation of structurally ambiguous sentences to save time, unless disambiguation is required by task demands as under deep processing. Committing to one reading and building the corresponding syntactic structure is argued to take time. Therefore, not having to make a commitment because of underspecification under shallow processing is expected to result in a reading time advantage for ambiguous sentences. The longer question answering times for ambiguous sentences under deep processing, as also found in our study, have been taken as additional evidence in favor of the delay of a commitment to a particular reading (Swets et al., 2008).

For referential ambiguities, Stewart et al. (2007) argued that ambiguous pronouns may initially be underspecified under shallow processing, resulting from the absence of co-indexation between the pronoun and its antecedent. According to this approach, readers invest processing resources to interpret the ambiguous pronoun only when they are required to engage in deep processing; under shallow processing ambiguous pronouns may remain underspecified. A sentence such as *Abel lent Gijs the book before he went on holiday* will be understood by the comprehender to mean that someone went on holiday, but it will be left undecided whether that was Abel or Gijs. Because of this underspecification, under shallow processing reading times for ambiguous sentences are expected to be faster or equivalent compared to unambiguous sentences. By contrast, under deep processing, reading times for ambiguous sentences are expected to be slower.

The finding of processing depth affecting ambiguous sentence processing can also be explained without the notion of underspecification. Logačev and Vasishth (2016a) proposed a Stochastic Multiple-Channel Model of structural ambiguity resolution as an extension of Van Gompel et al.'s (2000) Unrestricted Race Model. To account for the flexible nature of ambiguity resolution, Logačev and Vasishth (2016a) stipulated a *stopping rule* that is determined by task demands. If the task does not require more than one permissible relative clause attachment to be constructed, the parser stops after one attachment has been computed (i.e., a *first-terminating* stopping rule). However, if the task requires access to all available attachment options, the parser may choose to wait for all permissible structures to be built (i.e., an *exhaustive* stopping rule). Under a first-terminating stopping rule, an ambiguity *advantage* is predicted, because the parser needs to wait for only one particular attachment in unambiguous conditions and the fastest will be adopted. However, when the exhaustive stopping rule is used, an ambiguity *disadvantage* is predicted because the parser needs to wait for both attachments to be computed in the ambiguous condition, which will lead to more instances of long completion times.

This approach can easily be applied to referential ambiguities as well. Under deep processing, as implemented in the experiment reported here, the task suggests to the participant that an exhaustive stopping rule should be used, because the task requires resolving the pronominal references or because the reader is aware that the pronouns sometimes have more than one referent. In that case, the parser waits for both referents that form potential antecedents for the ambiguous pronoun to be computed (and then possibly discards the least plausible of them). This predicts an ambiguity penalty under deep processing, because computing both referents will always take longer than computing one, even in a parallel processing system (Logačev & Vasishth, 2016a). In contrast, under shallow processing conditions, one referent will suffice and an ambiguity advantage is predicted as the adopted referent is the one that takes the least time, leading to shorter time on average when there are more candidates as in the ambiguous cases.

In sum, both underspecification and race-based models are, in modified form, able to account for the finding that processing depth may influence the difference in reading times between referentially ambiguous and unambiguous sentences: the former through strategic underspecification of the co-indexation between the pronoun and its antecedent; the latter through the use of different stopping rules determined by task demands. A difference between the approaches is that at least one referent is selected as an antecedent to the pronoun in the Stochastic Multiple-Channel Model, while the co-indexation is not necessarily specified under an underspecification model. Hence, important questions are what the resulting representations look like under shallow processing and, under an underspecification approach, whether the resulting representations are partially specified or non-specified (see Logačev & Vasishth 2016b, for discussion). As pointed out by Van Gompel et al. (2001), the relevant parts of ambiguous sentences need to be processed to some extent under an underspecification approach as well, in order for the comprehension system to be able to determine what can be left underspecified (ambiguous cases) and what cannot (unambiguous cases). In response, Swets et al. (2008) argued that the relevant part is processed just enough to identify disambiguation cues and inappropriate interpretations, but not deeply enough to fully disambiguate when there are two appropriate interpretations. On the basis of the present data we cannot tell which representations were generated while the sentences were read and exactly how specified these were under shallow processing. It is possible that both referents remained available as possible antecedents, as predicted under an underspecification account. It is also possible that a specific referent was selected, either at random, or, more likely, because stored knowledge was taken into account (such as the bias for referents in subject position to be more accessible). Further research is needed to distinguish between these options and evaluate the validity of both models.

Thinking about the broader implications of the current findings, an important question is how representative our materials and task are for other reading/listening situations. With respect to the task, people do not usually face comprehension questions. Instead, depth of processing or the processing strategy more generally may be driven by both general constraints (e.g., whether the conversation or text is judged to be important) and local factors (e.g., information structure). The consequences of shallow/deep processing may also vary across situations. In our experiment, participants read isolated sentences in which the ambiguity always concerned a choice between two referents. Considering an underspecifiation model, these materials may have biased participants towards partial specification (e.g., *either Abel or Gijs went on holiday*), while a context in which the potential referents are not so clearly specified may bias participants towards complete underspecification (e.g., *someone went on holiday*).

Concluding remarks

In conclusion, the present results show, once again, that the way in which ambiguous sentences are processed depends on the depth of processing that participants engage in. This finding may not only be true for ambiguity resolution, but likely generalizes to other areas of language processing as well. For instance, recent work by Laurinavichyute and von der Malsburg (2021) showed that the illusion of ungrammaticality (i.e., a slow-down in reading due to agreement attraction in sentences like *The key to the cabinets* ***is*** *rusty*) disappeared when participants engaged in deep processing. In addition, a study by Schotter, Bicknell, Howard, Levy, and Rayner (2014) showed that the type of proofreading task that participants engaged in affected properties of word processing (such as word frequency and predictability effects). An important methodological implication of these findings is that the type of comprehension questions or tasks included in an experiment may affect the manner of processing. Care must, therefore, be taken in the choice of these comprehension questions and tasks (see also Ferreira & Yang, 2019). At a theoretical level, the present findings illustrate the need for models of ambiguity resolution that take into account the effect of processing depth on ambiguity resolution, and that can flexibly account for both ambiguity advantages and ambiguity penalties.

Data Accessibility Statement

The experimental stimuli, the data, and our analysis code are available at https://hdl.handle.net/1839/a3c0374b-8b05-4929-9458-a829f786b7ef.

Ethics and Consent

The experiments were conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from each participants prior to the start of the experiment. The study was approved by the ethics board of the Faculty of Social Sciences of Radboud University.

Acknowledgements

The authors would like to thank Thijs Rinsma for his assistance with programming the experiment, and Annelies van Wijngaarden, Carlijn van Herpt, and Dennis Joossen for their help with stimuli creation and participant recruitment. We also thank Laurel Brehm for useful discussions regarding the statistical analyses reported here, and three anonymous reviewers for comments on previous versions of this paper.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

AC and AM conceptualized the experiment, decided on the methodology, and contributed to the interpretation of the data. AC took the lead on data collection, data analysis, and writing; AM reviewed and edited.

References

Arnold, J. E. (1998). *Reference form and discourse patterns*. Stanford University.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*(1), B13–B26. DOI: https://doi.org/10.1016/S0010-0277(00)00073-1

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

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

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

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 817–828. DOI: https://doi.org/10.1080/17470218.2015.1134603

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407. DOI: https://doi.org/10.1006/cogp.2001.0752

Clifton, J. C., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass*, *2*(2), 234–250. DOI: https://doi.org/10.1111/j.1749-818X.2008.00055.x

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39* (2), 175–191. DOI: https://doi.org/10.3758/BF03193146

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15. DOI: https://doi.org/10.1111/1467-8721.00158

Ferreira, F., & Henderson, J. M. (1990). Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 555–568. DOI: https://doi.org/10.1037/0278-7393.16.4.555

Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good-enough language processing. In *Psychology of learning and motivation*, 65, 217–247. Elsevier. DOI: https://doi.org/10.1016/bs.plm.2016.04.002

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. DOI: https://doi.org/10.1111/j.1749-818X.2007.00007.x

Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes*, *56*(7), 485–495. DOI: https://doi.org/10.1080/0163853X.2019.1591885

Grant, M., Sloggett, S., & Dillon, B. (2020). Processing ambiguities in attachment and pronominal reference. *Glossa: A Journal of General Linguistics*, *5*(1), 77. DOI: https://doi.org/10.5334/gjgl.852

Greene, S. B., McKoon, G., & Ratcliff, R. (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 266. DOI: https://doi.org/10.1037/0278-7393.18.2.266

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes*, *26*(3), 376–405. DOI: https://doi.org/10.1080/01690965.2010.492642

Hofmeister, P., & Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology*, *5*, 1237. DOI: https://doi.org/10.3389/fpsyg.2014.01237

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. DOI: https://doi.org/10.1037/0096-3445.111.2.228

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040. DOI: https://doi.org/10.1080/17470218.2015.1053951

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. DOI: https://doi.org/10.3758/BRM.42.3.643

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmertest: Tests in linear mixed effects models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=lmerTest (R package version 2.0-32)

Laurinavichyute, A., & von der Malsburg, T. (2021). *Agreement attraction in grammatical sentences arises only in the good-enough processing mode.* Paper presented at the 34th Annual CUNY Conference on Human Sentence Processing, University of Pennsylvania.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R package version*, *1*(1), 3.

Logacev, P., & Vasishth, S. (2016a). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*(2), 266–298. DOI: https://doi.org/10.1111/cogs.12228

Logačev, P., & Vasishth, S. (2016b). Understanding underspecification: A comparison of two computational implementations. *Quarterly Journal of Experimental Psychology*, *69*(5), 996–1012. DOI: https://doi.org/10.1080/17470218.2015.1134602

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*(2), 157–201. DOI: https://doi.org/10.1080/01690969408402115

McDonald, J. L., & MacWhinney, B. (1995). The time course of anaphor resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language*, *34*(4), 543–566. DOI: https://doi.org/10.1006/jmla.1995.1025

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (https://www.R-project.org/)

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. *The Quarterly Journal of Experimental Psychology Section A*, *57*(5), 893–933. DOI: https://doi.org/10.1080/02724980343000549

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, *131*(1), 1–27. DOI: https://doi.org/10.1016/j.cognition.2013.11.018

Stewart, A. J., Holler, J., & Kidd, E. (2007). Shallow processing of ambiguous pronouns: Evidence for delay. *Quarterly Journal of Experimental Psychology*, *60*(12), 1680–1696. DOI: https://doi.org/10.1080/17470210601160807

Swets, B., Desmet, T., Clifton, J. C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216. DOI: https://doi.org/10.3758/MC.36.1.201

Traxler, M. J., Pickering, M. J., & Clifton, J. C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592. DOI: https://doi.org/10.1006/jmla.1998.2600

Van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *52*(2), 284–307. DOI: https://doi.org/10.1016/j.jml.2004.11.003

Van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In *Reading as a perceptual process*, 621–648. Elsevier. DOI: https://doi.org/10.1016/B978-008043642-5/50029-2

Van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, *45*(2), 225–258. DOI: https://doi.org/10.1006/jmla.2001.2773

Van Rij, J., Van Rijn, H., & Hendriks, P. (2013). How WM load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, *5*(3), 564–580. DOI: https://doi.org/10.1111/tops.12029

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4(ISBN 0-387-95457-0). DOI: https://doi.org/10.1007/978-0-387-21706-2_14