Understanding underspecification: A comparison of two computational implementations

Pavel Logačev and Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany

Version dated November 10, 2015

Abstract

Swets et al. (2008) present evidence that the so-called ambiguity advantage (Traxler et al., 1998), which has been explained in terms of the Unrestricted Race Model, can equally well be explained by assuming underspecification in ambiguous conditions driven by task-demands. Specifically, if comprehension questions require that ambiguities be resolved, the parser tends to make an attachment: when questions are about superficial aspects of the target sentence, readers tend to pursue an underspecification strategy. It is reasonable to assume that individual differences in strategy will play a significant role in the application of such strategies, so that studying average behavior may not be informative. In order to study the predictions of the good-enough processing theory, we implemented two versions of underspecification: the partial specification model (PSM), which is an implementation of the Swets et al. proposal, and a more parsimonious version, the non-specification model (NSM). We evaluate the relative fit of these two kinds of underspecification to Swets et al.'s data; as a baseline, we also fit three models that assume no underspecification. We find that a model without unspecification provides a somewhat better fit than both underspecification models, while the NSM model provides a better fit than the PSM. We interpret the results as lack of unambiguous evidence in favor of underspecification; however, given that there is considerable existing evidence for good-enough processing in the literature, it is reasonable to assume that some underspecification might occur. Under this assumption, the results can be interpreted as tentative evidence for NSM over PSM. More generally, our work provides a method for choosing between models of real-time processes in sentence comprehension that make qualitative predictions about the relationship between several dependent variables. We believe that sentence processing research will greatly benefit from a wider use of such methods.

Introduction

Sentence processing research has been focused on answering the question: How do we integrate the words we hear or read into syntactic structure in order to arrive at the

meaning of a sentence? Theories of sentence comprehension have typically assumed that readers or listeners create a fully specified representation of a sentence they are trying to understand. This means that in a sentence like (1), readers know that *the boy* is the agent and *the dog* is the patient of biting. It also means that in ambiguous sentences such as (2), readers either think that the general was standing on the balcony, or that the general's daughter was. In other words, a widely held assumption is that the comprehender attaches the relative clause to either the first noun (N1) or the second noun (N2).

(1) The boy bit the dog.

(2) Who saw the daughter$_{N1}$ of the general$_{N2}$ who was standing on the balcony?

However, there is increasing evidence that the relevant research question may well be: *do* we combine words to build structure at all? A prominent example is Christianson, Hollingworth, Halliwell, and Ferreira (2001); they found that readers sometimes do not carry out full reanalysis of garden-path sentences. In their experiment, participants read sentences such as (3b) and (3a). (3b) is a locally ambiguous version of the sentence in (3a), which tends to garden-path readers. When asked comprehension questions such as *Did the man hunt the deer?*, participants tended to respond 'yes' more often in the locally ambiguous condition (3b) than in the unambiguous baseline (3a). On the basis of findings such as these, Christianson et al. (2001) argue that participants do not always fully reanalyze garden-path sentences, and that they sometimes create an inconsistent representation of the sentence in (3b). In this representation, *the deer* functions as the object of *hunted*, but also as the subject of *ran*.

(3) a. While the man hunted the pheasant the deer ran into the woods.

b. While the man hunted the deer ran into the woods.

This finding is not unexpected under the assumptions of the good-enough approach to language comprehension (e.g., Ferreira, Bailey, & Ferraro, 2002; Sanford & Sturt, 2002). Under this view, the comprehender seeks to reduce processing effort, and tries to do no more than what they think is sufficient to complete the task. To this end, they may either underspecify certain aspects of the sentence meaning (Sanford & Sturt, 2002), or use heuristics to arrive at the final interpretation. For example, Ferreira (2003) found that participants were significantly worse at correctly identifying the patient and agent of implausible passive sentences like (4b) than their plausible counterparts such as (4a). There was no such difference between corresponding active sentences. According to Ferreira (2003), these findings suggest that readers may make use of simple heuristics instead of deploying their syntactic machinery when the latter would be too taxing.

(4) a. The man was bitten by the dog.

b. The dog was bitten by the man.

We thank Benjamin Swets for sharing the raw data from Swets, Desmet, Clifton, and Ferreira (2008), and for providing many helpful and constructive comments on our work, both in the present paper and in Logačev and Vasishth (2015).

In sum, proponents of the good-enough processing account have provided strong evidence that the comprehender does not always build perfect representations – instead they sometimes make use of simpler strategies which will produce the desired results, at least on some trials. Such a strategy may also be the explanation for the surprising finding called the *ambiguity advantage*, which we discuss next.

The Ambiguity Advantage

Traxler, Pickering, and Clifton (1998) found that ambiguous sentences like (5c) were read faster at the word *moustache* than their unambiguous counterparts such as (5a) and (5b). To explain this finding, Traxler et al. (1998) and van Gompel, Pickering, and Traxler (2000) proposed the *Unrestricted Race Model (URM)*. According to the URM, the parser commits to either an N1- or an N2-reading as soon as it encounters an ambiguity (i.e., at the word *that*). Importantly, whether the parser chooses to attach the RC to N1 or to N2 varies from trial to trial – in other words, the parser's choice is non-deterministic. As a result, the parser sometimes commits to parses in the unambiguous conditions, which later turn out to be wrong. Upon disambiguation at *moustache*, reanalysis is required. Ambiguous conditions, however, are compatible with either reading, and so no reanalysis is required. The lack of reanalysis leads to a speedup in the ambiguous conditions.

(5) a. The driver of the car *that had the moustache* was pretty cool. (N1 attachment)
b. The car of the driver *that had the moustache* was pretty cool. (N2 attachment)
c. The son of the driver *that had the moustache* was pretty cool. (globally ambiguous)

There are other interesting alternative explanations of the ambiguity advantage. In addition to a task-dependent variant of the unrestricted race model proposed by the present authors (Logačev & Vasishth, 2015), the ambiguity advantage can be explained by Levy's surprisal theory (Levy, 2008). Levy discusses the ambiguity advantage using the trio of sentences shown in (6).

(6) a. The daughter$_i$ of the colonel$_j$ who shot herself$_{i/*j}$ on the balcony had been very depressed. (N1 attachment)
b. The daughter$_i$ of the colonel$_j$ who shot himself$_{*i/j}$ on the balcony had been very depressed. (N2 attachment)
c. The son$_i$ of the colonel$_j$ who shot himself$_{i/j}$ on the balcony had been very depressed. (globally ambiguous)

Levy's surprisal account for the ambiguity advantage is that the conditional probability of the potentially disambiguating word (*himself* or *herself*) is higher in the ambiguous sentences like (6c) than in the unambiguous cases (6a) and (6b). This is because in (6c) both possible attachments of the RC, high and low, contribute probability mass to the probability of *himself* appearing (thus making it more predictable). In (6a), however, only N1 attachment contributes probability mass and in (6b) only the N2 attachment contributes probability mass (making *himself* less predictable in (6a) and (6b) than in (6c)). This is

certainly a possible explanation. However, in this article, our goal is to explore the implications of a radically different explanation of the ambiguity advantage, proposed by Swets et al. (2008).

Underspecification as an Explanation for the Ambiguity Advantage

Swets et al. (2008) claim that the ambiguity advantage is a consequence of *strategic underspecification*. According to their account, the comprehender underspecifies the meaning of ambiguous sentences if the task does not require ambiguity resolution. This behavior can explain the ambiguity advantage found by Traxler at al., because participants in that experiment did not have to answer questions about RC attachment. Therefore, ambiguity resolution was not required. An interesting prediction of the underspecification account is that the ambiguity advantage should disappear when the comprehender expects to be asked about relative clause attachment. Such expectations force the parser to disambiguate. Swets and colleagues tested this hypothesis in an experiment with sentences such as (7). While RC attachment (N1, N2, or ambiguous) was varied as a within-subject factor, three different groups of participants were asked different types of comprehension questions. Forty-eight participants were asked questions concerning RC attachment after every experimental sentence (e.g., *Did the maid/princess/son scratch in public?*). Another group of 48 participants was asked superficial questions which were unrelated to RC attachment (e.g., *Was anyone humiliated/proud?*), and a third group was asked superficial questions occasionally.

(7) a. The son of the princess who scratched himself in public was terribly humiliated. (N1 attachment)

b. The son of the princess who scratched herself in public was terribly humiliated. (N2 attachment)

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

Swets et al. found an ambiguity advantage in the superficial questions condition, but not in the RC questions conditions. This finding is consistent with the predictions of strategic underspecification: In the superficial condition, the comprehender does not expect to be tested about RC attachment and thus underspecifies to conserve time and effort. In the RC questions condition, however, this is not an option. Consequently, the comprehender attaches the relative clause to the preferred attachment site (i.e., to N2) most of the time.

In addition to effects of question type on reading time, Swets et al. found that participants were slower at answering RC attachment questions about ambiguous sentences than about unambiguous sentences. They argue that the additional time needed to answer RC questions when the sentence is ambiguous can be explained by assuming that the parser sometimes underspecifies RC attachment during reading, and that the RC has to be attached before responding to the question. This postponed RC attachment is carried out after reading the question and requires additional time. This implies that even in the RC questions condition, underspecification trials occur. To put it differently, even in conditions where participants are expected to carry out the attachment, in some cases they do not make an attachment.

We have previously argued elsewhere (Logačev & Vasishth, 2015) that the finding of slowed responses in ambiguous conditions does not constitute clear-cut evidence in favor of underspecification. This is, in part, because the underspecification model makes a more fine-grained prediction: it predicts that the subset of trials affected by underspecification should be associated with longer question answering times, as well as faster reading. However, Swets et al. did not find an ambiguity advantage in reading times in the RC questions condition. Importantly, the absence of this finding is in principle compatible with the underspecification model, under the assumption that underspecification trials cause a sufficiently large slow-down during the question answering phase, but a relatively small speed-up during reading. Because this explanation assumes a mixture of trials which are not straightforwardly separable (underspecification and non-underspecification), this hypothesis can best be tested by directly modeling this mixture, which we will do in this paper.

Furthermore, if underspecification occurs even when task-demands presumably require full specification, a fuller treatment of underspecification in the good-enough framework needs to answer the following questions: (a) what exactly happens during underspecification trials; (b) how often does underspecification occur? In order to spell out the logical possibilities, we formalize the Swets et al. model of underspecification, develop a more parsimonious version of this model, and compare these two models' relative fit with respect to the Swets et al. data.

Before we can present the alternative models of underspecification, it is important to understand the salient facts of the Swets et al. study first. We address this point below.

A Reanalysis of Swets et al.'s Data

We analyzed the response accuracy, the question-answering time, and the reading time data from the RC questions condition in Swets et al.'s experiment.[1] We only analyzed data from the RC questions condition, because the three dependent measures pertinent to underspecification were recorded on each trial: reading time, question answering time, and the RC attachment indicated by the response (N1 attachment, or N2 attachment). Because across all question conditions, Swets et al. (2008) found effects of attachment on the potentially disambiguating word (*himself/herself*) and the spill-over region (*in public*), we analyzed the time participants required to read both regions (treated as one region). In the analysis, we used all trials with question answering times of less than 15 seconds.[2] We excluded the data of 11 out of 48 participants because they had 50% or more errors in answering questions about one of the unambiguous conditions. Of the excluded participants, 5 were excluded due to errors in the N1 attachment condition, and 6 due to errors in the N2 attachment condition. We excluded their data because such high error percentages may be indicative of a reading strategy in which readers consistently attach either to N1 or to N2, irrespective of the evidence provided. Such reading strategies, although potentially interesting and worth further study, may also indicate that these participants may have pursued a reading strategy which is outside the scope of the present work.

Table 1 shows the average reading time at the critical region, *himself/herself in public*. It shows that the non-local N1 attachment conditions are read more slowly than the

[1] Many thanks to Benjamin Swets for providing us with the raw data of the experiment.

[2] However all the patterns reported here held true when we applied a stricter exclusion criterion of 8 seconds.

ambiguous and local N2 attachment conditions. This could be either because (a) the parser always attempts the local N2 attachment first, even in N1 attachment conditions; or (b) N1 attachment is slower because the first noun, which is more distant from the relative clause, requires more time to be retrieved from memory than the second noun. Although ambiguous sentences are read somewhat more slowly than N2 attachment sentences, the difference is not significant (the 95% confidence interval for reading times in the ambiguous conditions is $[1832\,ms; 2084\,ms]$). Table 2 shows the average question response time by attachment condition. Question-responses in ambiguous conditions are slower than in unambiguous conditions, and questions about N2 attachment sentences are answered faster than questions about N1 attachment sentences.

Table 1
Mean reading times (in milliseconds) for the critical region, by attachment. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

N1 attachment	N2 attachment	ambiguous
2143 (63)	1845 (44)	1958 (63)

Table 2
Mean question answering times (in milliseconds) for RC questions, by attachment. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

N1 attachment	N2 attachment	ambiguous
2826 (98)	2512 (86)	3033 (116)

Table 3 shows the average proportions of responses indicating N2 attachment by attachment condition. For example, a 'yes'-response to a N1 question is considered to indicate N1 attachment, while a 'no'-response is considered to indicate N2 attachment. While participants answered questions about unambiguous sentences with an accuracy of approximately 80%, the percentage of responses indicating N2 attachment in ambiguous sentences was closer to 60%, suggesting that the preference for N2 attachment was relatively weak.

Table 3
Mean proportions of responses indicating N2 attachment by attachment condition. Standard errors in brackets.

N1 attachment	N2 attachment	ambiguous
0.22 (0.02)	0.83 (0.02)	0.59 (0.02)

Table 4 shows the average reading times in unambiguous conditions at the critical region as a function of response correctness. It shows that reading times for trials associated with incorrect responses tend to be numerically shorter for N1 attachment sentences than those associated with correct responses. For N2 attachment sentences, the pattern is reversed. However, neither difference is statistically significant.

Table 5 shows the average question-answering time as a function of the correctness of the answer to the comprehension question. It shows that participants take more time to respond incorrectly than correctly. A possible reason is that they first try to retrieve

Table 4
Mean reading times (in milliseconds) in the unambiguous condition at the critical region by correctness of the response. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

	N1	N2
correct response	2165 (70)	1834 (47)
incorrect response	2064 (94)	1902 (81)

the memory trace of the sentence representation, fail at doing so, and then initiate a guess. Whatever the correct explanation for the delay, it points towards an interpretation that incorrect responses stem from a qualitatively different process requiring more time than is required for an ordinary response.

Table 5
Mean question-answering times in unambiguous conditions by attachment and correctness of the response. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

	N1	N2
correct response	2641 (98)	2382 (84)
incorrect response	3489 (216)	3172 (214)

To summarize the insights from Tables 1-5:

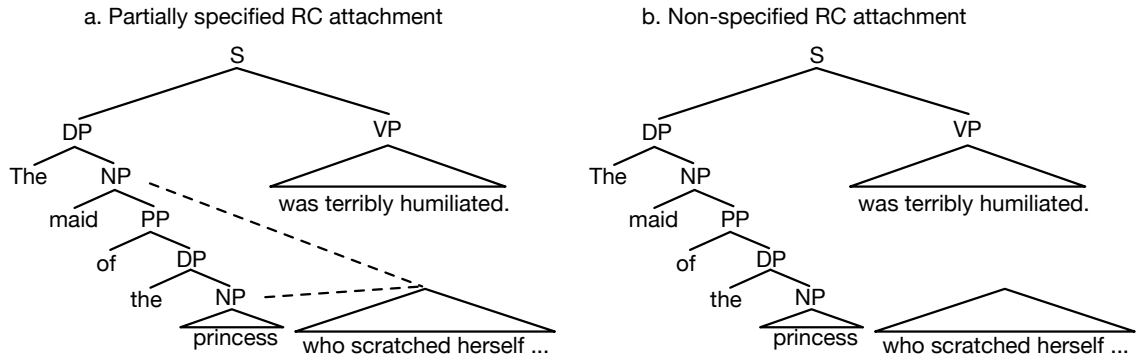
1. At the critical region, non-local N1 attachment conditions are read more slowly than local N2 attachment and ambiguous conditions.
2. Question-response times in ambiguous conditions are slower than in unambiguous conditions, and questions about N2 attachment sentences are answered faster than questions about N1 attachment sentences.
3. In ambiguous sentences, the proportion of responses consistent with an N2 attachment was approximately 60%, suggesting a weak preference for N2 attachment in the face of global ambiguity.
4. In unambiguous sentences, there was no statistically significant effect of response correctness on reading times at the critical region. However, there was a tendency towards shorter reading times for N1 attachment sentences followed by incorrect responses, and towards longer reading times for N2 attachment sentences followed by incorrect responses.
5. Question-response times were longer for incorrect responses to unambiguous sentences, compared to response times for correct responses.

We discuss next the implications of these facts for the underspecification account of Swets et al.

Partial Specification and Non-Specification

Swets et al. claim that readers are able to attach the RC during question-answering on underspecification trials. This claim entails that the parser must remember which noun phrases are potential attachment sites—if this information were absent, the reader would

Figure 1. Two kinds of underspecified representations: partially specified (left), and non-specified (right).



have to either re-parse the sentence completely, or examine each noun phrase in memory as a potential attachee, a potentially very expensive operation. Thus, the parser must *store* information about potential attachment sites even when it underspecifies. As a result, we must assume that the underspecified representation of sentence (7c) looks like the one shown in figure 1a. We will refer to this kind of underspecification as a *partial specification*, because partial information about RC attachment is stored by the parser. We will not assume that information about attachment is stored in a particular format. Therefore, partial specification is consistent with underspecified representations such as those proposed by Frazier and Clifton (1997) or by Sturt and Crocker (1997). What is important is that information about RC attachment *is* stored.

Importantly, the ambiguity advantage found by Traxler et al. and in the superficial questions conditions of the Swets et al. experiment is not straightforwardly compatible with partial specification. This is because the parser needs to store attachment-related information in ambiguous as well as unambiguous conditions. Therefore, underspecification will be predicted to be faster than RC attachment only if we stipulate that creating a partial specification requires less time than completing the attachment (i.e., fully specifying the attachment).[3] This may well be a reasonable assumption; but prima facie, establishing a memory for a potential attachment site (and of the co-dependents to be attached) could take just as much time (or more) as actually completing the dependency.

However, partial specification is not the only possible way to implement underspecification. An alternative explanation for the ambiguity advantage (the speedup in ambiguous sentences) is that the parser does not save any information at all about potential attachment sites in the ambiguous condition; this is a departure from the assumption that Swets et al. must make, as discussed above. Figure 1b illustrates the resulting structure of sentence (7c). The parser keeps information about the main clause and about the relative clause, but it does not associate the RC with any of the noun phrases. The difference between

[3] A possible explanation for why partial specification requires less time than full unambiguous specification is that ambiguous attachments are not semantically interpreted and that establishing one syntactic link *and* semantically interpreting it requires more time than establishing two syntactic links. However, this explanation, too, requires stipulations about the relative durations of processes.

partial specification and what we will refer to as *non-specification* of RC attachment is that in non-specification, potential attachment sites are not marked as such. Thus, in order to save time, the parser does not do anything attachment-related, and this results in an ambiguity advantage.

An obvious drawback of not storing attachment information is that no attachment can be carried out after reading the comprehension question, at least not without a prohibitively expensive reparsing process. Therefore, in trials where the comprehender engages in non-specification, they have to resort to guessing the answer to the question.[4] If we assume that guessing requires more time than informed question-answering, we can explain why relative clause questions are answered more slowly when they are about ambiguous sentences than when they are about unambiguous sentences. This assumption, that guessing consumes more time than informed question-answering, is consistent with the pattern in table 5, which shows longer response times in incorrect responses. As discussed above, these longer RTs may represent a failed attempt to retrieve the syntactic representation, followed by a guess; if the total guessing time subsumes these two steps, it seems reasonable to assume that guessing takes longer than an informed decision. Importantly, non-specification is more parsimonious than partial specification to the extent that they can account for the data equally well, because the latter needs to stipulate that partial specification requires less time than full specification, whereas the non-specification model does not require such stipulations.

What are the consequences of these two alternative theories of underspecification? A computational implementation has the potential to shed light on this question. We describe next the implementation details of the partial specification and non-specification models.

A Model of Partial Specification

According to Swets et al.'s proposal, the reading time and question answering data in the ambiguous condition must consist of a mixture of trials. Figure 2a shows a flow diagram of the partial specification model which implements the logic of Swets et al.'s account of the results in the RC questions condition.

The figure shows that when attachment is unambiguous, readers have only one option: attaching the relative clause.[5] The processor can choose the attachment site either based on syntactic information (with probability $1 - p_{NSYN}$), or based on non-syntactic information (with probability p_{NSYN}). The latter option is motivated by the finding that, in some situations, the processor may choose to ignore syntactic cues in the processing of unambiguous sentences and base its interpretation on processing heuristics instead (Ferreira, 2003; Christianson, Luke, & Ferreira, 2010). Furthermore, this assumption is in line with the pattern in table 4, according to which reading times on trials with incorrectly answered compre-

[4] A further prediction of the non-specification hypothesis is that on non-specification trials in sentences like (1), no information is kept on whether the RC can attach to *the general*, *the assistant*, or *the CEO*.

(1) Mary showed the general the assistant of the CEO who was standing on the balcony.

[5] While it is possible to interpret Swets et al.' proposal such that underspecification occurs in both, ambiguous and unambigous sentences, but more frequently in ambiguous sentences, it is not clear why this would be so. Therefore, we will adopt the simplifying assumption that underspecification affects only ambiguous sentences.

hension questions tend to be lower than on trials with correctly answered questions for N1 sentences (i.e., closer to reading times in the N2 attachment condition), while the opposite is true for N2 sentences (i.e., reading times on trials with incorrectly answered questions are higher and thus closer to reading times for N1 attachment sentences). This suggests that on a proportion of those trials, readers might be creating the wrong attachment, resulting in longer (or shorter) reading times than on correctly processed trials. Importantly, while *syntactically driven RC attachment* always results in a correct sentence interpretation, *non-syntactically driven RC attachment* can result in correct or incorrect interpretations, depending on whether the cues used in determining the attachment site are aligned with the sentence structure.

Because non-syntactic interpretations may disagree with the sentence structure, this alternative route of interpretation can explain some of the incorrect responses in table 3. However, it fails to account for the finding that incorrect responses are slower than correct responses (cf. table 5). Because both, the syntactic and the non-syntactic route, result in RC attachment, the time required to answer questions about it should be the same in both cases.

A further potential explanation is that although readers process unambiguous sentences in the regular manner, they sometimes fail to retrieve the representation of the relevant part of the sentence (with probability p_{FAIL}) during the question-answering process. As a result, they try to guess the correct answer. Importantly, we will assume that guessing requires more time than regular question-answering, as discussed above in connection with the pattern in table 5.[6] Because the retrieval-failure explanation and the non-syntactic RC attachment explanation make different predictions about question-response latencies, we will be able to assess the relative importantce of non-syntactic RC attachment and of guesses in the explanation of incorrect responses.

Because non-syntactic RC attachment does not make use of syntactic cues, attachment sites must be chosen based on the same information as in ambiguous conditions (e.g., thematic information, or linear-order-based heuristics). Therefore, we assume that disambiguation in ambiguous conditions is carried out by the same process as non-syntactic attachment. Because such a process may take into account the linear order of potential attachment sites, it may be exhibit an N2 attachment preference (e.g., Carreiras & Clifton, 1993). This is in line with the findings presented in table 3, according to which there is a bias towards N2 responses in the ambiguous condition. To capture this bias in the model, we assume that non-syntactic RC attachment results in N2 attachment with probability $1 - p_{N1}$, and in N1 attachment with probability p_{N1}, where p_{N1} is a free parameter to be estimated from the data.

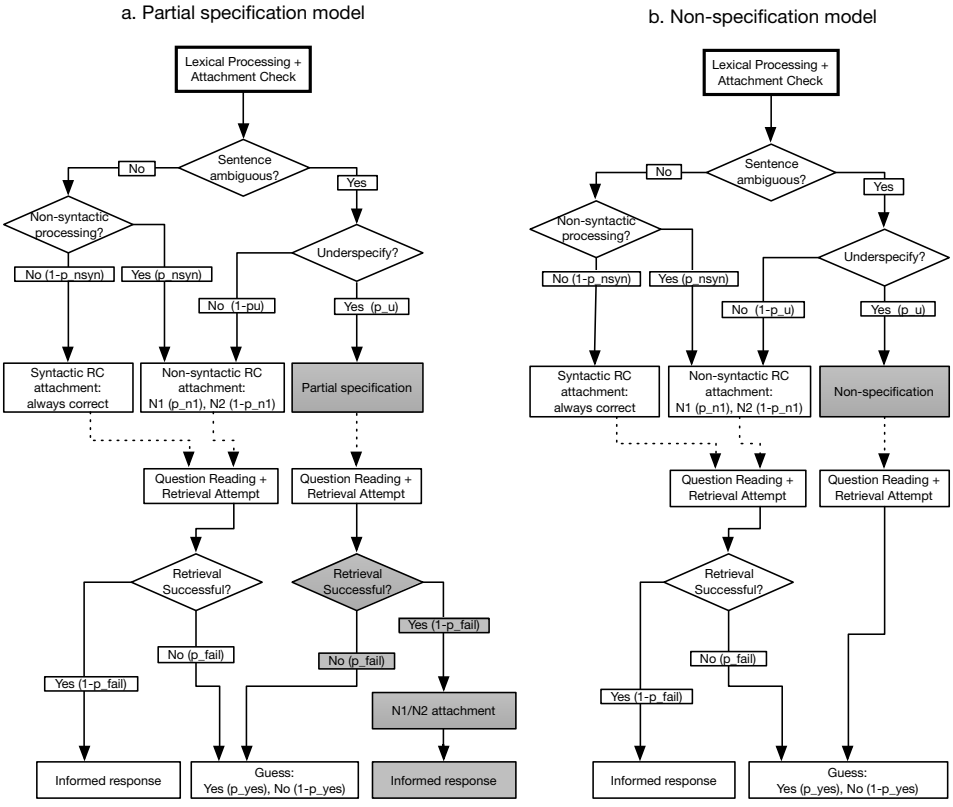
According to the partial specification model in figure 2a, participants underspecify

[6]There are several explanations for why guesses could be slower than informed decisions: When readers fail to recall the sentence structure, they start searching their memory for clues as to the correct answer. This extensive search in memory could slow down the response. Alternatively, responding to a question could involve a competition between the two response options in a manner similar to the competition-integration model (Spivey & Tanenhaus, 1998; Spivey-Knowlton, 1996). Lack of evidence either way could prolong competition, and thus cause long response times when information about RC attachment is either not present or cannot be recalled. Yet another possibility is uncertainty about past input (Levy, Bicknell, Slattery, & Rayner, 2009); in this case, an incorrect representation of the previously processed material could lead to a retrieval failure due to a retrieval cue not matching the intended target.

the structure of ambiguous sentences on some trials (with probability p_U). The critical assumption of the partial specification model is that on those trials, only a partially specified structure, as in fig. 1a, are generated. Because creating a partially specified representation requires less time than creating a full representation, reading is fast on underspecification trials. However, questions are answered slowly, because the previously omitted RC attachment needs to be carried out during the question-answering stage, before responding. On non-underspecification trials on the other hand, RC attachment is carried out during reading which is why participants read more slowly, but are faster at answering questions. Irrespective of whether a full or a partially specified sentence representation is stored, its retrieval can fail during the question-answering phase; in such a case, a guess as to the correct answer is generated.

We will also assume, in agreement with the results in table 3, that the probability of retrieval failure does not depend on the attachment condition or the attachment-related operation carried out during reading (attachment, or underspecification).

Figure 2. Flow-charts of the sequence of operations according to the the partial specification model (left panel), and according to the non-specification model (right panel). Probabilities of decisions are shown in brackets where appropriate. Differences between the two models are highlighted in grey.



To summarize, the partial specification model makes the following assumptions about

the parser's operations:

1. Readers always fully specify RC attachment in unambiguous sentences, either using syntactic information (with probability $1 - p_{NSYN}$), or using non-syntactic heuristics (with probability p_{NSYN}). In the latter case, they choose to attach the RC to N2 (with probability $1 - p_{N1}$) or to N1 (with probability p_{N1}).
2. In ambiguous sentences, readers may choose to underspecify with probability p_U. When readers do not underspecify the attachment (with probability $1 - p_U$), they attach the RC during reading using non-syntactic information to determine the attachment site.
3. Answering questions about RC attachment requires the retrieval of (parts of) the sentence representation. Retrieval of the corresponding memory trace may fail with probability p_{FAIL}, irrespective of whether it is fully or partially specified.
4. When retrieval fails, comprehenders attempt to guess the answer. They answer 'yes' with probability p_{YES}, and 'no' with probability $1 - p_{YES}$. Guessing requires more time than giving an informed response.
5. The regular mechanism required for question-answering can only operate on fully specified representations, and so readers attempt to disambiguate partially specified representations before answering a question. However, disambiguation can only take place if an underspecified representation is successfully retrieved.

6a. When the parser underspecifies, it stores information about potential attachment sites, thus allowing for postponed RC attachment if necessary. We call this assumption 6a because the alternative proposal (presented below), will make a different assumption (6b).

In addition, our implementation of the partial specification model makes the following assumptions about the timing of processes:

1. Partial specification requires less time than full specification of N1 or N2 attachment.
2. N2 attachment requires less time than N1 attachment. This assumption is motivated by the findings presented in table 1.
3. Generating a guess as to the correct response to a question requires more time than giving an informed response. This assumption is motivated by the findings presented in table 5.

A Model of Non-Specification

Like the partial specification model, the non-specification model posits a mixture of different kinds of trials as well. In the unambiguous conditions, reading is assumed to proceed as in the partial specification model: participants always carry out RC attachment, using syntactic or non-syntactic cues, as illustrated in figure 2b. In most cases, this attachment is followed by an informed response to the comprehension question, but in some cases, participants have to guess the answer due to a failed retrieval from memory.

When reading ambiguous sentences, readers can either attach the RC or choose to underspecify attachment, just like in the partial specification model. When the RC is attached, comprehenders proceed like in the partial specification model. They give informed

responses to comprehension questions, but sometimes, when retrieval of the sentence representation fails, they have to resort to guessing.

The crucial difference between the two models lies in what constitutes underspecification during reading. While the partial specification model assumes that some information about potential attachment sites is stored (as in fig. 1a), the non-specification model assumes that no such information is stored (as in fig. 1b). A consequence of the latter assumption is that the non-specification parser cannot choose to fully specify RC attachment at a later point.

Thus, the key difference in the predictions of the two models is that according to non-specification, question responses on underspecification trials consist of guesses only, whereas according to the partial specification model they consist of (i) some guesses and (ii) some informed responses preceded by RC attachment during the question answering phase. However, both accounts agree that underspecification trials are preceded by faster reading.

In sum, the non-specification model makes the same assumptions about the timing of processes as the partial specification model, as well as assumptions 1-5 about the parser's operations. Instead of assumption 6a, however, the non-specification model adopts assumption 6b.

6b. When the parser underspecifies, it does *not* store any information about potential attachment sites, and thus does not allow for postponed RC attachment. As a result, the only way to answer a question on underspecification trials is to guess.

Implementation of the Two Underspecification Accounts, and of Non-underspecification Models as Baselines

The partial specification model and the non-specification can both explain Swets et al.'s finding that questions are answered more slowly when they are about ambiguous sentences than when they are about unambiguous sentences. Both models predict that this slowdown in question-answering is caused by underspecification trials, i.e., trials on which RC attachment is underspecified. Importantly however, the models make different predictions about the quantitative relationship between reading time, question-response time, and response patterns on underspecification trials.

The partial specification model predicts that response times on underspecification trials are longer than on non-specification trials by the amount of time required to attach the RC. This means, for example, that the difference in reading times between underspecification trials and N2 attachment trials should be equal to the difference in question-response times between N2 attachment trials and underspecification trials.[7] Furthermore, underspecification trials followed by postponed N1 or N2 attachment should result in responses indicating such attachment in most cases.

[7] An alternative possibility is that RC attachment requires more time when it is carried out during question-answering than when it is carried out during reading. Although it is to be expected that retrieval of the sentence representation will take more time during the question-answering phase than during reading, retrieval is involved in the answering of questions about unambiguous sentences as well. Thus, longer attachment times during question-answering can only be caused by a slowdown in the RC attachment operation *after* the sentence representation has been retrieved. However, it is not clear what could cause such a slowdown.

The non-specification model, on the other hand, predicts that response times on underspecification trials should be equal to the time required to generate a guess, i.e., they should be equal to response times for erroneous responses in unambiguous conditions. Furthermore, such responses should indicate N1 and N2 attachment with equal probability.

Because these predictions cannot be tested without obtaining estimates of RC attachment durations, as well as of the proportion of underspecification trials and their response latencies, we formalized both models in Stan (Stan Development Team, 2015) in order to estimate the model parameters and formally compare the quantitative fits of the models to the data. The simultaneous estimation of model parameters and model comparison will allow us to answer the following questions: Firstly, to what extent do the data agree with Swets et al.'s explanation for the slow responses to questions in the ambiguous condition? In other words, is there evidence for a subset of underspecification trials involving faster reading and slower responses than non-underspecification trials? Secondly, is the response time pattern on underspecification trials closer to the predictions of the partial specification model or to those of the non-specification model? Thirdly, how often do readers underspecify?

In order to compare the underspecification models to baseline no-underspecification models, we also implemented three versions of a model assuming no underspecification, i.e., models in which the probability of underspecification is 0. These three no-underspecification models assume (i) retrieval failure and non-syntactic RC attachment, (ii) no retrieval failure, and (iii) no non-syntactic RC attachment.

Method

We implemented hierarchical versions both models according to the flow charts in fig. 1 in Stan (Stan Development Team, 2015). We assumed that all reading times and reaction times follow a gamma distribution (e.g., Luce, 1986). We estimated one common parameter for (1) the scale and separate shape parameters for (2) base reading time per word (i.e., underspecification), (3) N1 attachment, (4) N2 attachment, (5) informed question answering, and (6) guessing. Furthermore, we estimated (7) the probability of failing to recall a sentence during question-answering (p_{FAIL}), (8) the probability of underspecifying in the ambiguous condition (p_U), (9) the probability of choosing the attachment site based on non-syntactic information (p_{NSYN}), (10) the probability of choosing an N1 attachment when attaching the RC in the ambiguous condition (p_{N1}), and (11) the probability of guessing 'yes' (p_{YES}). We assumed that by-participant parameter values for all probability parameters followed a beta distribution (the prior on the probabilities was a vague $Beta(1, 1)$ distribution), and that all other parameters were distributed log-normally. We treated the grand mean and the variance of the by-participant parameters as free parameters to be estimated as part of the model.

As mentioned above, in order to determine the relative importance of non-syntactic RC attachment and retrieval failure in accounting for incorrect responses, we first fit three models without any underspecification ($p_U = 0$): we compared a model assuming retrieval failure and non-syntactic RC attachment (p_{FAIL} and p_{NSYN} were free parameters) to a model assuming no retrieval failure ($p_{FAIL} = 0$), and to a model assuming no non-syntactic RC attachment ($p_{NSYN} = 0$). In a second step, in order to quantify the evidence in favor of underspecification, we compared the models with no underspecification to a partial specification and a non-specification model. Model comparison was done on the basis of

the WAIC (Watanabe-Akaike Information Criterion; Watanabe, 2010; Vehtari & Gelman, 2014), which is an estimate of the model generalizability and rewards better fit to the data (lower WAIC) while penalizing model flexibility (higher WAIC). For each of the five models, we ran 4 chains with 2000 iterations.

Results and Discussion

Table 6 shows the WAIC and parameter estimates along with Bayesian *credible intervals* for the three models without underspecification. The credible interval can be interpreted as the range which contains the true parameter value with 95% probability (e.g., Lynch, 2007). The first column of the table shows that in order to account for incorrect responses without assuming retrieval failure, one has to assume that participants ignore syntactic cues on approximately 40% of the trials ($\hat{p}_{nsyn} = 39\%$, CI: $31\% - 47\%$), and that they choose to attach to N1 in approximately 40% of those cases ($\hat{p}_{N1} = 41\%$, CI: $35\% - 47\%$). This finding agrees with an error rate of approximately 20% in table 3: in many cases, readers form a correct interpretation of the sentence in spite of ignoring syntactic cues, resulting in only 20% incorrect responses. Table 6 also shows that the model accounting for erroneous responses without retrieval failures needs to assume substantially higher response times for informed responses ($2647ms$) than the other two models ($2043ms$ and $2123ms$, respectively) can generate only informed responses. Therefore, the average response time for informed responses is just the average response time.

Importantly, table 6 also shows that the WAICs are substantially higher for the model assuming no retrieval failure ($WAIC = 44821.0$) and for the model assuming no non-syntactic RC attachment ($WAIC = 44532.0$) than for the model assuming both ($WAIC = 44248.1$). This means that response times associated with incorrect responses are best described by the model assuming both mechanisms, i.e., as mixture of two distributions: relatively fast responses due to incorrect RC attachment when syntactic cues are ignored ($Est. = 2123ms$, CI: $2073 - 2175$), and relatively slow responses due to guessing ($Est. = 5129ms$, CI: $4954 - 5305$).

Fig. 3 (left panels) shows the predictions of the best model without underspecification alongside the average response percentages indicating N1 attachment, as well as reading and response times from the Swets et al. data: while it can account for the patterns in question-answering and reading times, it predicts equal response times in all attachment conditions. This is because this model does not implement any underspecification ($p_U = 0$) and thus assumes equal proportions of guesses and infomed responses in all conditions.

Table 7 shows the parameter estimates for both underspecification models. The credible intervals for p_U show that the estimated proportion of underspecification trials is relatively low for both models: $0 - 17\%$ of all trials for the the partial specification model, and $1 - 12\%$ for the non-specification model.

The remaining parameter estimates do not substantially differ between the two underspecification models, and do not deviate by much from the estimates of the maximal model without underspecification in table 6. The WAIC slightly favors the non-specification model ($WAIC = 44259.6$) over the partial specification model ($WAIC = 44263.1$), which is likely due to the fact that the non-specification model can better account for the increased response time in the ambiguous conditions, as shown in fig. 3 (bottom). However, the difference in WAICs ($\Delta WAIC = 3.5$) is relatively small given in relation to the standard

Table 6
Models with no underspecification: Parameter estimates and WAIC. (95% credible intervals in brackets.)

	No underspecification ($p_{FAIL} = 0$)	No underspecification ($p_{NSYN} = 0$)	No underspecification (all parameters)
p_{NSYN}	.39 [.31-.47]	–	.28 [.20-.35]
p_{FAIL}	–	.39 [.33-.44]	.23 [.19-.28]
p_{N1}	.41 [.35-.47]	.33 [.24-.42]	.38 [.32-.44]
p_{YES}	.51 [.06-.92]	.45 [.40-.51]	.46 [.39-.51]
READING TIME: N2 ATTACHMENT	1902 [1851-1953]	1845 [1801-1891]	1821 [1780-1863]
READING TIME: N1 ATTACHMENT	2320 [2264-2383]	2205 [2151-2262]	2169 [2119-2210]
RESPONSE TIME: INFORMED	2647 [2594-2704]	2043 [1980-2109]	2123 [2073-2175]
RESPONSE TIME: GUESS	–	4089 [3948-4231]	5129 [4954-5305]
$WAIC$	44821.0 (SE=150.0)	44532.0 (SE=143.2)	44248.1 (SE=131.6)

error for the differences in WAIC ($SE_{\Delta WAIC} = 5.7$), and so the evidence in favor of the non-specification model given the present data can be considered very weak at best.

Importantly, both underspecification models had higher WAIC values than the maximal model without underspecification in table 6 ($WAIC = 44248.1$), indicating that the latter is more likely to generalize well to future data. Because these differences, too, are relatively small in comparison to the standard error of the WAIC differences ($\Delta WAIC = 11.5, SE_{\Delta WAIC} = 10.4$ and $\Delta WAIC = 15.0, SE_{\Delta WAIC} = 9.4$ respectively), they do not allow us to rule in favor of one or another model. However, this finding suggests that there is little evidence for the existence of underspecification trials, i.e., trials with faster-than-normal reading, and slower-than-normal question answering. This is also why both underspecification models predict shorter response times in the ambiguous condition than were actually observed, as can be seen in fig. 3. In order to predict higher response times in that condition, one needs to assume a higher percentage of underspecification trials. However, doing so would cause the model to underpredict the average reading time in ambiguous conditions. Thus, the predictions in fig. 3 represent the optimal point in the trade-off between prediction error for reading times and reaction times.

In sum, we found evidence for two types of trials which can lead to incorrect responses: trials with relatively fast responses, preceded by non-syntactic RC attachment, and trials with relatively slow responses due to guessing, following a failure to retrieve. Furthermore, we found very little evidence for the existence of underspecification trials. Higher WAIC values for underspecification models and estimates indicating a low probability of underspecification indicate that if readers underspecify, they do so very rarely.

Table 7
Underspecification models: Parameter estimates and WAIC. (95% credible intervals in brackets.)

	Partial specification	Non-specification
p_U	.07 [.00-.17]	.06 [.01-.12]
p_{NSYN}	.27 [.19-.35]	.27 [.20-.35]
p_{FAIL}	.22 [.18-.27]	.22 [.17-.26]
p_{N1}	.39 [.31-.45]	.39 [.32-.45]
p_{YES}	.45 [.38-.53]	.46 [.39-.54]
READING TIME: UNDESPECIFICATION	1338 [1231-1439]	1345 [1231-1453]
READING TIME: N2 ATTACHMENT	1841 [1797-1888]	1843 [1793-1912]
READING TIME: N1 ATTACHMENT	2174 [2124-2231]	2170 [2114-2223]
RESPONSE TIME: INFORMED	2119 [2062-2174]	2123 [2065-2179]
RESPONSE TIME: GUESS	5140 [4935-5364]	5101 [4893-5318]
$WAIC$	44263.1 (SE=131.6)	44259.6 (SE=131.7)

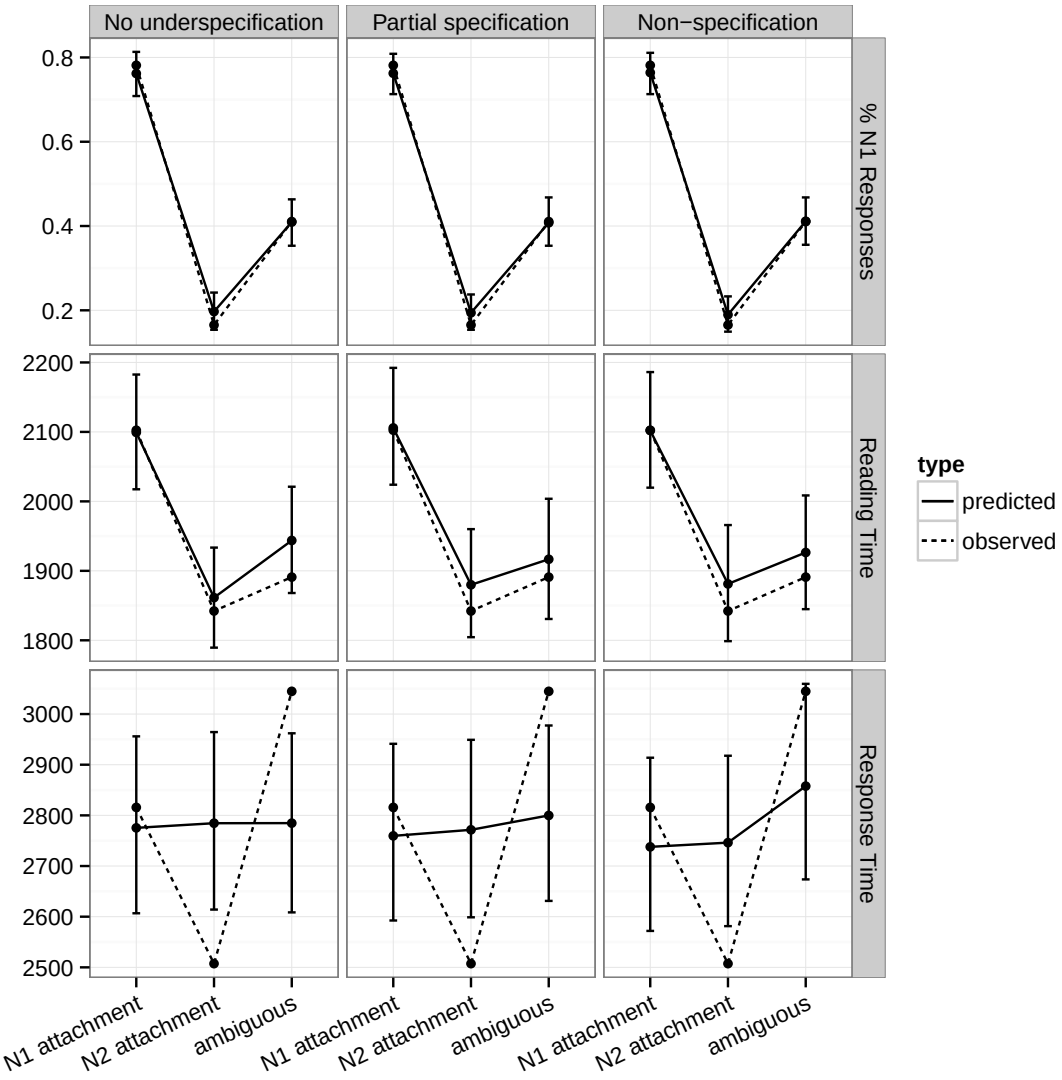
General Discussion

We have presented two models which are compatible with Swets et al.'s finding of longer question-answering times following ambiguous conditions than unambiguous conditions. The *partial specification model* is an extension of Swets et al.'s original underspecification model and is based on the assumption that ambiguous sentences can be underspecified, but that some information about potential RC attachment sites is stored. This means that RCs with an initially underspecified attachment can be accessed and attached at a later point in time. According to this model, answering questions about ambiguous RC attachment should require more time because RC attachment is carried out during the question-answering phase.

We further proposed an alternative, the *non-specification model*, which also bears the reading time signature of underspecification, i.e., it predicts fast reading in ambiguous sentences, followed by slower question-answering than in unambiguous sentences. This model assumes that no information about potential attachment sites is stored. Therefore, no RC attachment can take place during the question answering phase because the parser does not know which noun phrases in memory are viable candidates for attachment. Thus, participants try to guess the right answer, which requires more time than providing an informed response on trials where RC attachment took place during reading.

We argued that because both models posit a mixture of trials which are not straightforwardly separable, their quality of fit is best assessed by directly modeling this mixture. We then estimated the model parameters for several candidate models and compared the models' relative quality of fit. In order to obtain correct estimates of process durations, we first compared three models without underspecification. We found evidence for the as-

Figure 3. Predictions of (i) the model without underspecification (left), (ii) the partial specification model (center), and (iii) the non-specification model (right) in comparison to the results from Swets et al.'s experiment. The plots shows percentages of N1 responses (upper panels), reading times (central panels), and question-answering latencies (lower panels). Error bars on the model predictions correspond to 95% credible intervals.



sumption that readers sometimes ignore syntactic cues in RC attachment, as well as for the assumption that, in spite of having made an attachment during reading, readers sometimes fail to retrieve the processed sentence structure and have to resort to guessing the answer to a comprehension question. We therefore incorporated both of these assumptions into our implementation of the underspecification models.

We found that both underspecification models can account for the data nearly equally

well, with the non-specification model providing a slightly better quantitative account of the longer question-response times in the ambiguous condition. Furthermore, the estimates of both underspecification models show that underspecification, to the extent that it exists, affects less than 17% of all trials in the ambiguous conditions. This low percentage is in line with our finding that both underspecification models had a somewhat worse fit according to their WAICs than a model assuming no underspecification: A potentially slightly better fit for the underspecification models was offset by the additional model flexibilitly due to the additional parameters related to the probability of underspecification.

In conclusion, the models investigated suggest that underspecification, to the extent that underspecification is an adequate account of the ambiguity advantage, appears to be a relatively rare phenomenon given the Swets et al. data. Because of its low frequency of occurrence in these data, both models provide an equally good account of underspecification. A major achievement of the present work is that we develop a methodology for formalizing and testing the assumptions underlying underspecification, and we evaluate the empirical evidence for different instantiations of underspecification. Although it is highly plausible that some form of underspecification is in play in day-to-day language use, the range of underspecification strategies deployed in sentence comprehension needs closer investigation, using carefully controlled experiments such as Swets and colleagues', and computational modeling.

Conclusion

We have presented two different models of underspecification and tested their predictions on the data of Swets et al. (2008). We found evidence that two different mechanisms are responsible for incorrect responses, one of which is likely to reflect readers' failure to employ syntactic cues during processing, while the other appears to reflect failure to retrieve the sentence meaning followed by guessing. We further found that in the data we investigated there is very little evidence for the existence of underspecification trials as proposed by Swets et al. (2008) (i.e., faster reading of the relative clause, followed by slowed response to questions). We also found that under the assumption that readers do underspecify, they underspecify only rarely. Thus, the empirical evidence from controlled studies for underspecification remains an open question; it is also unclear at present what exactly triggers underspecification in some trials but not in others. However, we are confident that these issues can be addressed in future research by investigating experimental data through the lens of computational modeling, as we have done here. Moreover, we have demonstrated a method for choosing between models of real-time processes in sentence comprehension that make qualitative predictions about the relationship between several dependent variables. We believe that sentence processing research will greatly benefit from a wider use of such methods.

References

Carreiras, M., & Clifton, C. (1993). Relative clause interpretation preferences in spanish and english. *Language and Speech*, *36*(4), 353–372.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic Roles Assigned along the Garden Path Linger. *Cognitive Psychology*, *42*(4), 368–407.

Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 538.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 2–5.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Frazier, L., & Clifton, C. J. (1997). Construal: Overview, motivation, and some new evidence. *Journal of Psycholinguistic Research*, *26*(3), 277–295.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Logačev, P., & Vasishth, S. (2015). A Multiple-Channel Model of Task-Dependent Ambiguity Resolution in Sentence Comprehension. *Cognitive Science*.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Lynch, S. M. (2007). *Introduction to applied bayesian statistics and estimation for social scientists*. Springer Science & Business Media.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61–64.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521.

Spivey-Knowlton, M. K. (1996). *Integration of visual and linguistic information in spoken language comprehension* (Doctoral dissertation). Unversity of Rochester.

Stan Development Team. (2015). Stan: a c++ library for probability and sampling, version 2.7.0.

Sturt, P., & Crocker, M. (1997). Thematic monotonicity. *Journal of Psycholinguistic Research*, *26*(3).

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, *36*(1), 201–216.

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592.

van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process.* Oxford: Elsevier.

Vehtari, A., & Gelman, A. (2014). Waic and cross-validation in stan. *Submitted. http://www.stat.columbia.edu/˜gelman/research/unpublished/waic_stan. pdf*.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.