UNDERSPECIFICATION AND PARALLEL PROCESSING IN SENTENCE COMPREHENSION

by

Pavel Logačev

Doctoral Thesis submitted to the Faculty of Human Sciences at the University of Potsdam (Department Linguistik) in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Department Linguistik,
University of Potsdam
submitted, June 2014

First reviewer: Prof. Dr. Shravan Vasishth
Second reviewer: Prof. Dr. Kiel Christianson
Day of oral defense: December 09, 2014

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-82047
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-82047

ABSTRACT

The aim of the present thesis is to answer the question to what degree the processes involved in sentence comprehension are sensitive to task demands. A central phenomenon in this regard is the so-called ambiguity advantage, which is the finding that ambiguous sentences can be easier to process than unambiguous sentences. This finding may appear counterintuitive, because more meanings should be associated with a higher computational effort. Currently, two theories exist that can explain this finding.

The *Unrestricted Race Model* (URM) by van Gompel et al. (2001) assumes that several sentence interpretations are computed in parallel, whenever possible, and that the first interpretation to be computed is assigned to the sentence. Because the duration of each structure-building process varies from trial to trial, the parallelism in structure-building predicts that ambiguous sentences should be processed faster. This is because when two structures are permissible, the chances that *some interpretation* will be computed quickly are higher than when only one specific structure is permissible. Importantly, the URM is not sensitive to task demands such as the type of comprehension questions being asked.

A radically different proposal is the *strategic underspecification* model by Swets et al. (2008). It assumes that readers do not attempt to resolve ambiguities unless it is absolutely necessary. In other words, they *underspecify*. According the strategic underspecification hypothesis, all attested replications of the ambiguity advantage are due to the fact that in those experiments, readers were not required to fully understand the sentence.

In this thesis, these two models of the parser's actions at choice-points in the sentence are presented and evaluated. First, it is argued that the Swets et al.'s (2008) evidence against the URM and in favor of underspecification is inconclusive. Next, the precise predictions of the URM as well as the underspecification model are refined. Subsequently, a self-paced reading experiment involving the attachment of pre-nominal relative clauses in Turkish is presented, which provides evidence against strategical underspecification. A further experiment is presented which investigated relative clause attachment in German using the speed-accuracy tradeoff (SAT) paradigm. The experiment provides evidence against strategic underspecification and in favor of the URM. Furthermore the results of the experiment are used to argue that human sentence comprehension is fallible, and that theories of parsing should be able to account for that fact. Finally, a third experiment is presented, which provides evidence for the sensitivity to task demands in the treatment of ambiguities. Because this finding is incompatible with the URM, and because the

strategic underspecification model has been ruled out, a new model of ambiguity resolution is proposed: the *stochastic multiple-channel model* of ambiguity resolution (SMCM). It is further shown that the quantitative predictions of the SMCM are in agreement with experimental data.

In conclusion, it is argued that the human sentence comprehension system is parallel and fallible, and that it is sensitive to task-demands.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to Professor Shravan Vasishth for his incredible support throughout the entire duration of my PhD, as well as before that. Shravan introduced me to psycholinguistics as well as research methods and ignited my interest in both. He was always extraordinarily encouraging concerning whichever projects I chose to pursue. Equally importantly, he was always available for discussion and showed great interest in improving the quality of my research as well as furthering my career as a researcher.

The present thesis would not have been possible without his continuous advice and support, which is why I will abstain from using the first-person singular pronoun throughout this thesis.

Further, I would like to thank my academic colleagues and friends Samar Husain, Felix Engelmann, Titus von der Malsburg, Lena Jäger, Andreas Peldszus, Lena Benz, and Paul Metzner for the countless insights they shared with me, for their support, their encouragement, and for making research even more fun.

I especially would like to thank Umesh Patil, Rukshin Shaher, Bruno Nicenboim and Andreas Pankau for what would take pages to enumerate, if I were going to do so, but above all for being good friends, as well as for not hesitating to tell me when I am wrong.

The contribution of my partner and good friend, Özlem, cannot possibly be overstated. She was instrumental in every aspect of the thesis, from sanity-checking my ideas in the early research stages down to proof-reading the manuscript. Thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

"One morning, I shot an elephant in my pajamas.
How he got into my pajamas I'll never know."
Groucho Marx, in Animal Crackers (1930)

One of the most intriguing properties of human language is the ubiquitous presence of ambiguity. Although it is a source of amusement to many, as in the above quote, it also presents a challenge to theories of human sentence comprehension. After reading the first sentence of the above quote, for example, the reader assumes that it was the speaker who was wearing pajamas and not the elephant. Thus, the disambiguation in the second sentence is unexpected. But how exactly do we arrive at this first interpretation? While it is tempting to assume that we base our interpretations solely on plausibility, there is evidence to the contrary. For example, the comprehension difficulty caused by sentences such as (1) (Bever, 1970) suggests that structural factors play a major role in real-time sentence comprehension.

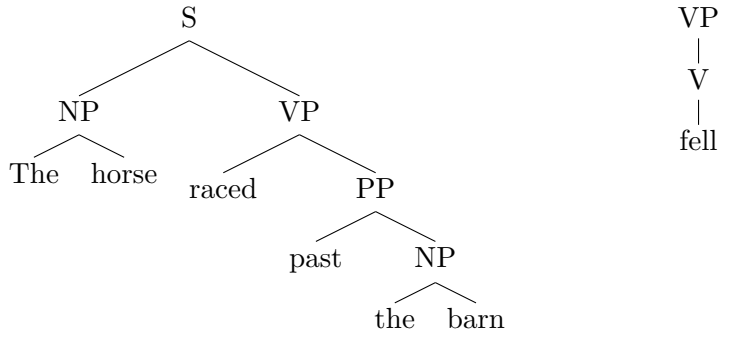
Introspection with regard to the source of the processing difficulty in (1) suggests that we might understand the first part of the sentence as a main clause such as (2). In effect, we fail to integrate the verb *fell*, as illustrated in figure 1.1. In other words, we experience a *garden-path effect*.

(1) The horse raced past the barn fell.

(2) The horse raced past the barn.

(3) The horse ridden past the barn fell.

Figure 1.1: Incorrect analysis of *The horse raced past the barn fell.*



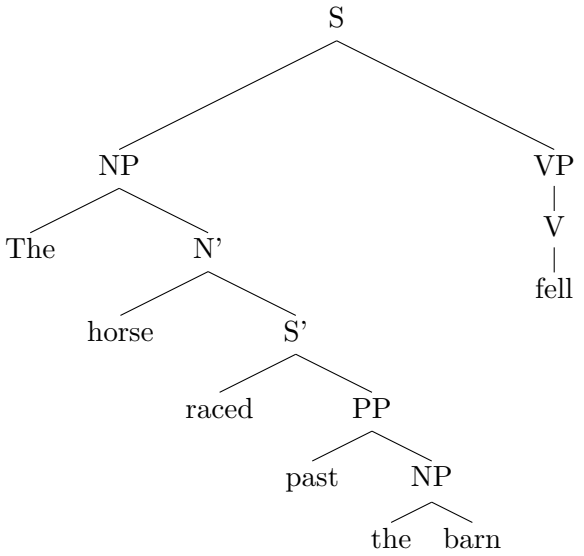
The sentence in (3), on the other hand, causes no processing difficulty because *The horse ridden past the barn* cannot be interpreted as a main clause, since *ridden* is a past participle. The verb form *raced* (1), however, is ambiguous between a past participle functioning as the verb of the reduced relative clause *(that was) ridden past the barn*, and the main verb of the sentence, such as in (2).

The fact that sentences such as (1) cause processing difficulty while sentences such as (3) do not, suggests that when the human sentence comprehension system (hereafter, *the parser*) is faced with a choice between two options, it sometimes tends to make somewhat premature decisions. For example, when faced with the choice of analyzing *raced* in (1) as a main clause verb, or as the beginning of a reduced relative clause, the parser chooses the main clause interpretation. This interpretation later turns out to be incorrect results in the syntactically incoherent structure in figure 1.1). The parser should have chosen the past participle interpretation, which results in the structure in figure 1.2. The result of the wrong choice is in sentence (1) is serious processing difficulty — a garden-path effect.

Because ambiguity is ubiquitous in our daily communication, and because in spite of the presence of garden-path effects, we appear to be able to deal with it very seamlessly most of the time, any theory of human sentence has to address the question of how the human sentence comprehension mechanism handles choice points when building structure incrementally.

An early and influential idea is the Garden-Path Model (e.g., Frazier, 1979; Frazier & Rayner, 1982), which assumes that the sentence comprehension system builds a detailed syntactic representation as one reads or hears a sentence. Because sentences need to be understood quickly, the parser's operation is guided by principles which minimize structure building cost. For example, according to the *Minimal Attachment Principle*, it attempts to build the least complex structure compatible with the

Figure 1.2: Correct analysis of *The horse raced past the barn fell.*



input so far. This principle explains the garden-path effect for sentences such as (1): According to Minimal Attachment, the parser constructs the structure in figure 1.1 because doing so requires less time than building the structure in figure 1.2, because the latter consists of more syntactic nodes. According to the Garden-Path Model, the parser's initial decisions are based only on syntactic information, which means that it operates deterministically — given a particular type of syntactic ambiguity, the parser always makes the same attachment decision. It is also serial, which means that, if a choice exists, only one syntactic structure is adopted.

Over time, all these claims—seriality, determinism, complete structure building, and the priority of syntax—have come to be challenged.

Constraint-based models (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998) abandon the syntax-first assumption, allowing all kinds of information such as syntax, semantics, and plausibility to be used simultaneously for making parsing decisions. For example, in McRae et al.'s (1998) implementation of the Competition-Integration Model, all permissible analyses of a sentence compete for activation. In this process, each analysis receives support from several sources (*constraints*). For example, in the sentence fragment (4), competition starts when the parser reaches the word *sent*. It is faced with the choice of analyzing *sent* either as a main verb, or as a past participle. If it chooses the main verb reading, the sentence can continue as in (5a). If it chooses the past participle reading, it must assume that sentence structure is that of reduced relative clause as in (5b), where *who was* was omitted. Sentence

(5b) has the same structure as the garden-path sentence in (1).

(4) The florist sent ...

(5) a. The florist sent the flowers was very pleased.
b. The florist *(who was)* sent the flowers was very pleased.

In the fragment in (4), the main clause reading (MC) of the verb is supported by the facts that (i) main clauses occur more frequently than reduced relative clauses, and (ii) the verb *sent* occurs more frequently as a main verb than as a past participle. The reduced relative clause reading (RR), on the other hand, is supported by the fact that (iii) florists are more likely to send flowers than to receive them because the former is a part of their job. Such sources of support for one reading over another are referred to as *constraints*.

According to the Competition-Integration Model, the graded support from several such constraints is translated into activation for each of the two readings (MC and RR). Subsequently, the two analyses compete for activation. The duration of the competition process is assumed to depend on the amount of evidence in favor of each of the analyses. If one reading is strongly preferred, the competition ends quickly. If however, both readings have approximately equal amounts of activation, the parser requires more time to choose one of them.

While the Competition-Integration Model can successfully model effects of plausibility and other constraints on reading times in sentences like (5b) (McRae et al., 1998), its core prediction is that competition only occurs during the reading of ambiguous sentences. Therefore, unambiguous sentences should be processed faster than ambiguous sentences due to the lack of competition.

Traxler, Pickering, and Clifton (1998) tested this prediction with sentences such as 6. In sentence (6c), the attachment of the relative clause *that had the moustache* is ambiguous because it can attach to *the driver* as well as to *the son*. In sentences (6a) and (6b), on the other hand, attachment is unambiguously *high* (first noun) or *low* (second noun), because a car cannot possibly have a moustache. In an eye tracking experiment, Traxler et al. (1998) found that in sentences like (6), the potentially disambiguating word (*moustache*) was read faster in the ambiguous condition (6c) than in the unambiguous conditions (6a) and (6b). Traxler et al. argued that this so-called *ambiguity advantage* is incompatible with constraint-based theories of sentence comprehension, because instead of a speed-up in ambiguous sentences

compared to their unambiguous counterparts, constraint-based theories predict a slowdown.

(6) a. The driver of the car *that had the moustache* was pretty cool. (high attachment)
b. The car of the driver *that had the moustache* was pretty cool. (low attachment)
c. The son of the driver *that had the moustache* was pretty cool. (globally ambiguous)

The surprising finding that ambiguity appears to facilitate reading has since been replicated by van Gompel, Pickering, and Traxler (2001), van Gompel, Pickering, Pearson, and Liversedge (2005), and Swets, Desmet, Clifton, and Ferreira (2008), and thus appears to be a fairly robust effect in reading. It poses a significant challenge to current theories of sentence comprehension, because it cannot be reconciled with deterministic theories of sentence processing that assume that the parser always behaves in same manner, no matter the task demands. Presently, two mechanistic models can account for this effect: the *unrestricted race model* (URM) by van Gompel, Pickering, and Traxler (2000) assumes that disambiguation is inherently non-deterministic and that the parser's decisions are influenced by random noise. The *strategic underspecification account* by Swets et al. (2008) assumes that sentence comprehension is inherently goal-directed and that readers are in control of their depth-of-processing during reading. This means that they are able to simply ignore certain aspects of the sentence meaning if they do not think that they will become relevant later. This behavior, too, is assumed to be non-deterministic. A further model can account for the ambiguity advantage: According to the *surprisal theory* (Levy, 2008), ambiguous sentences like (6a) are read faster than their locally ambiguous counterparts like (6b) and (6c) because the conditional probability of the potentially disambiguating word given the context is higher in ambiguous sentences. This is so because it can occur in both readings of the ambiguous sentence, but only in one reading of the unambiguous sentences. However, the surprisal theory is non-mechanistic in that, while it does provide a metric which quantifies processing difficulty (probability), it does not posit an algorithm such as a sequence of parsing operations which leads to the speed-up in the ambiguous conditions. Therefore, contrasting this theory with the previous two accounts is not straightforward and beyond the scope of the present thesis.

The present work is centered around the unrestricted race model and the strate-

gic underspecification account of the ambiguity advantage. Chapter 2 presents and discusses the unrestricted race model (URM), and strategic underspecification, as well as the Swets et al.'s (2008) evidence for the latter model. Chapter 3 investigates Swets et al.'s claims in more detail by renalyzing Swets et al.'s data and by presenting a version of the URM which is consistent with Swets et al.'s findings. It concludes that there is no evidence against the URM and in favor of the underspecification model. At this point, both models appear to be tenable explanations of the ambiguity advantage. Chapter 4 attempts to sharpen the vague notion of underspecification by presenting two quantitative models of underspecification, and comparing their relative fits to Swets et al.'s data. Although both models appear to be able to account for the data equally well, their parameter estimates suggest that underspecification may be deterministic — each participant either always underspecifies the relative clause attachment in ambiguous conditions or never underspecifies it. Chapter 5 presents the results of a self-paced reading experiment in Turkish, a language with pre-nominal relative clauses. The URM and the underspecification model make diverging predictions for Turkish. While the underspecification model predicts an ambiguity advantage in Turkish, the URM predicts no such effect. We find no ambiguity advantage in Turkish, a finding which is consistent with the URM, but not with the underspecification model. Chapter 6 uses the response signal paradigm to test the predictions of the URM and the strategic underspecification model in an experiment with German relative clauses and finds that firstly, ambiguous sentences are not only processed faster in some situations, but also more *successfully*. Secondly, it argues that the URM provides a more parsimonious account of our findings than underspecification. Chapter 7 presents the stochastic multiple-channel model (SMCM), an extension of the URM which is sensitive to task demands. It points out shortcomings of Swets et al.'s method for demonstrating the parser's ability to adapt to task demands, and presents the results of a German self-paced reading experiment, which are in line with the predictions of the SMCM. Quantitative predictions for the reading time data are derived from the SMCM, which seem to agree with the data. Chapter 8 concludes with an overview of the present work and argues that, presently, there is no evidence for strategic underspecification.

All computational modeling in the following chapters will be based on the simplifying assumptions that the parser (i) uses only syntactic cues in building sentence structure, and that (ii) it is purely reactive, i.e., that it does not engage in prediction. Although there is ample evidence for the influence of non-syntactic cues on parsing decisions (e.g. Ferreira, 2003; Christianson, Luke, & Ferreira, 2010; McRae et al., 1998), as well as for the role of prediction in parsing (e.g., Hale, 2001; Levy, 2008;

Altmann & Mirković, 2009), the integration of corresponding components into the computational models presented in the following is beyond the scope of this thesis.

Chapter 2

Two theories of ambiguity resolution

The present work will focus on Traxler et al.'s (1998) finding that readers tend to read ambiguous sentences faster than their unambiguous counterparts. This chapter will lay the groundwork for the discussion concerning the explanation of this finding in subsequent chapters. Two theories of how humans handle choice-points in sentence comprehension — the unrestricted race model and strategic underspecification — will be presented, as well as experimental evidence by Swets et al. (2008) which aims to distinguish between them empirically.

In subsequent chapters, we will review the assumptions of both models as well as the evidence in favor of each in more detail. We will further attempt to distinguish between them empirically, and arrive at the conclusion that the URM offers the most parsimonious account of the empirical findings so far.

We will discuss the unrestricted race model next.

2.1 The Unrestricted Race Model (URM)

Recall that Traxler et al. (1998) found in an eye tracking experiment that ambiguous sentences such as (6c, repeated as 7c) were read faster than their unambiguous counterparts (6a, repeated as 7a) and (6b, repeated as 7b). This speed-up was found at the potentially disambiguating word moustache.

To account for this effect, van Gompel et al. (2000) proposed a new model of ambiguity resolution, the *Unrestricted Race Model* (URM). According to the URM, when the parser encounters an ambiguity, it starts building all permissible structures simultaneously.

The time taken to construct a particular structure depends not only on structure's syntactic complexity, but also on other properties of the corresponding interpretation, including plausibility. Moreover, that time is assumed to vary as a function of random noise in the construction process.

Since the adopted reading in each trial is the one that takes the least time to be constructed, the structure-building process is *non-deterministic* due to the influence of random noise. This means that, on a given trial, any one of the readings can be adopted, because any of the structure-building processes could finish faster and thus win the race. Importantly, parsing is assumed to be strictly incremental in the sense that all structure-building is carried out at the earliest possible point. This means that in all of the sentences in (6) (repeated as (7)), the parser attaches the relative clause *that had the moustache* as soon as it encounters the relativizer *that*.

(7) a. The driver of the car *that had the moustache* was pretty cool. (high attachment)

b. The car of the driver *that had the moustache* was pretty cool. (low attachment)

c. The son of the driver *that had the moustache* was pretty cool. (globally ambiguous)

This means that on some trials, ambiguous sentences such as (7c) will receive a high attachment interpretation, which means that the relative clause (RC) will be attached to first noun (*son*). On other trials they will receive a low attachment interpretation, which means that the RC will attach to the second noun (*driver*). In (7a) and (7b), the chosen attachment turns out to be wrong on some trials and the sentence has to be reanalyzed as soon as the parser encounters *moustache*. In the ambiguous sentence in (7c), however, reanalysis is never necessary because it is compatible with any attachment. Thus, the locus of the ambiguity advantage is at *moustache*, the point of disambiguation in (7a) and (7b).

In sum, the URM is a non-deterministic model of disambiguation, because it assumes that a variety of factors, including random noise, contribute towards disambiguation.

All of the above-mentioned models — the Garden-Path Model, the Competition-Integration Model, as well as the URM — focus on answering one question: how do we decide which syntactic structure to assign to the words we hear or read in order to arrive at the meaning of a sentence? However, there is evidence that the relevant research question may well be: *do* we combine words to build structure at all?

2.2 Strategic Underspecification

2.2.1 Good-enough Processing

Most theories of sentence comprehension assume that readers or listeners create a fully specified representation of the sentence that they are trying to understand. This means that in a sentence like (8), readers know that *the boy* is the agent and *the dog* is the patient of biting. It also means that in globally ambiguous sentences such as (9), 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 (*high attachment*) or to the second noun (*low attachment*).

(8) The boy bit the dog.

(9) Who saw the daughter of the general who was standing on the balcony?

However, readers might not be creating fully specified representations of sentences at all times. 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 (10a) and (10b). (10b) is a locally ambiguous version of the sentence in (10a), which tends to garden-path readers. During the reading of such sentences, *the deer* is typically first analyzed as the object of *hunted*. Once the verb *ran* is encountered, *the deer* has to be reanalyzed as the subject of the main clause.

When participants were asked comprehension questions such as *Did the man hunt the deer?* about the sentences (10a) and (10b), they tended to respond 'yes' more often in the locally ambiguous condition (10b) than in the unambiguous baseline (10a). 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 (10b). In this representation, *the deer* functions as the object of *hunted*, but also as the subject of *ran*.

(10) 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 attempts 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 a plausible interpretation. For example, Ferreira (2003) found that participants were significantly worse at correctly identifying the patient and agent of implausible passive sentences like (11b) than of their plausible counterparts such as (11a). There was no such difference between corresponding active sentences. According to Ferreira (2003), these findings suggest that because passives are more difficult to understand than active sentences, readers may make use of simple heuristics instead of deploying their syntactic machinery when the latter would be too taxing.

(11) a. The man was bitten by the dog.
b. The dog was bitten by the man.

In sum, proponents of the good-enough processing account have provided evidence that comprehenders do not always build perfect representations – instead they may try to make use of simpler strategies which will produce the desired results, at least some of the time. Such a strategy may also be the explanation for the *ambiguity advantage*. This will be discussed next.

2.2.2 Strategic Underspecification

Swets, Desmet, Clifton, and Ferreira (2008) (SDCF, henceforth) suggest an alternative to the URM, which is grounded in the good-enough approach to sentence comprehension. Their explanation is based on the observation that in the studies concerning the ambiguity advantage (Traxler et al., 1998; van Gompel et al., 2001, 2005), reading comprehension was ensured by occasional superficial questions, which were intended to not draw attention to the attachment ambiguity. According to SDCF, such task demands did not require the parser to resolve the ambiguity in these studies, unless explicit disambiguation was provided, as in (7a) and (7b). Hence, the attachment was expected to remain underspecified in the globally ambiguous conditions.

This is presumably so because, if an ambiguity is detected, not making any commitment is less costly (when the task doesn't require it) than committing to one reading and building the corresponding structure.

Since participants in the above-mentioned studies were never asked questions about

the relative clause attachment after reading an experimental sentence, strategic underspecification could be a feasible strategy for reducing processing effort.

To test this explanation, Swets and colleagues asked participants to read sentences like (12) and asked different kinds of questions about them. The difficulty and frequency of the questions was manipulated in a between-participants design. While 48 participants were asked questions about relative clause attachment on every experimental trial (e.g., *Did the maid/princess/son scratch in public?*), another group of 48 participants was asked superficial questions (e.g., *Was anyone humiliated/proud?*). A further group of 48 participants was asked superficial questions only occasionally (once every 12 trials).

An ambiguity advantage was found when questions were superficial. However, consistent with the underspecification hypothesis of SCDF, no such effect could be found when all questions were about the attachment of the relative clause. When such questions were asked, the globally ambiguous condition (12c) was read as fast as the low attachment condition (12b), while the high attachment condition (12a) was read more slowly.

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

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

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

Swets and colleagues argue that the URM is unable to explain these data, because it predicts that globally ambiguous sentences should be processed faster than locally ambiguous sentences, irrespective of the kinds of questions asked. They explain the task dependence of the ambiguity advantage in terms of strategic underspecification: if questions are simple and do not require disambiguation of the sentence, the parser does not try to commit to any particular reading unless provided with a clear disambiguation cue. If the task does require disambiguation, however, the ambiguity is resolved towards the preferred reading. In this case, the parser will need the same amount of time for ambiguous sentences and for those disambiguated towards the preferred reading.

In addition, Swets and colleagues found that questions concerning RC attachment were answered more slowly after ambiguous sentences than after their unambiguous

counterparts. They interpret this finding as additional evidence for underspecification, because according to them, the parser sometimes (but rarely) underspecifies ambiguous sentences even when questions are about the RC. When this happens, the ambiguity has to be resolved before answering the comprehension questions, and this additional operation slows down processing. Unfortunately, Swets et al. do not discuss whether these two different behaviors (underspecification and full structure-building) are due to an inherent non-determinism of the parser, or possibly due to some participants not paying attention to the task.

2.3 Summary

In this chapter, two models accounting for the ambiguity advantage were presented: the unrestricted race model (van Gompel et al., 2000), and strategic (Swets et al., 2008). Furthermore, experimental evidence against the URM was presented. The main findings were (i) that the presence of the ambiguity advantage appears to depend on the task to be performed in the experiment, and (ii) that RC questions are answered more slowly when they are about ambiguous sentences, than when they are about unambiguous sentences. SDCF argue that both findings are incompatible with the URM, because it (i) it is not susceptible to task demands, and (ii) it always creates fully specified representations of sentences. In the next chapter, these arguments will be discussed in turn. It will be shown that while the argument based on question-answering latencies does not stand up to scrutiny, the URM can be reconciled with the finding of the task dependence of the ambiguity advantage.

Chapter 3

A re-examination of the Evidence against the URM

Swets et al. (2008) argued against the URM on the basis of two findings: (i) RC questions were answered more slowly when they were about ambiguous sentences than when they were about unambiguous sentences, but no such difference was observed for superficial questions. (ii) An ambiguity advantage was observed when questions were superficial, but no such effect occurred in the RC questions condition.

It seems unclear how the URM could explain either of these findings. It cannot predict a difference in question-answering times because it assumes that sentence structures are always fully specified. Furthermore, it appears unable to account for the lack of an ambiguity advantage in the RC questions condition because it is not sensitive to task demands. The validity of these two arguments will be discussed in turn.

3.1 Lack of Evidence for Underspecification from Question-response Times: Reanalysis of Swets et al.'s Question-response Latencies

Swets and colleagues found that question-answering times were significantly longer after ambiguous than after unambiguous sentences when RC questions were asked. No such differences were found when questions were superficial. SDCF interpret this finding as additional evidence for underspecification, because, according to them, participants *sometimes* (but rarely) underspecify ambiguous sentences even when

questions target the RC. In such cases, RC attachment is not carried out until a question about RC attachment has to be answered. Their argument rests on the assumption that the content of the question cannot be compared to an underspecified representation. Therefore, the reader cannot arrive at an answer to the question without carrying out attachment first. Thus, answering questions about ambiguous sentences should require more time than about unambiguous sentences because carrying out attachment during the question-answering phase (*question-triggered RC attachment*, henceforth) requires additional time, as compared to cases where a fully specified representation exists and no attachment needs to be carried out, either because sentences are unambiguous or because questions are superficial.

The strategic underspecification account predicts longer question answering times after ambiguous sentences than after unambiguous sentences, but only when questions target the RC. This is because in unambiguous sentences, RC attachment is carried out during reading. Their model also predicts no such difference when questions are superficial. As a consequence, the underspecification model predicts an interaction between the factors *question type* and *attachment*.

However, SDCF did not test for an interaction in question answering times, because *"the questions were so vastly different in the two question conditions"* (Swets et al., 2008, p. 209). Since it is exactly that difference between questions that is assumed to drive the effect, we think it is justified to test for an interaction between the two factors. Only if the effect of attachment is significantly larger in the RC questions condition can we consider longer question-answering times in the RC questions condition to be evidence for strategic underspecification of ambiguous sentences.

3.1.1 Method

We reanalyzed SDCF's question-response times[1] with linear mixed-effects models (Pinheiro & Bates, 2000; Baayen, 2008; Gelman & Hill, 2007) using *lme4.0* package (Bolker, Maechler, Bates, & Walker, 2013) in R (R Core Team, 2013). We included fixed effects of attachment and question type into the linear mixed-effects model, as well as random intercepts for participants and items. We did not include random slopes for attachment or question type, since the simulations presented by Barr, Levy, Scheepers, and Tily's (2013) suggest that their non-inclusion does not decrease statistical power.

We used treatment contrasts for the factor *question type* with superficial questions

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

coded as 0 and RC questions as 1. For the factor *attachment*, we used sliding contrast coding (e.g., Venables & Ripley, 2002) with comparisons between *ambiguous vs. high*, and *high vs. low*. We compared ambiguous to high attachment conditions instead of comparing them to the average of high and low attachment conditions because strategic underspecification predicts that answers to questions should be slowest when sentences are ambiguous. Therefore, they should be slower than the slowest unambiguous condition, which is the high attachment condition.

We conducted all analyses on log-transformed reaction times, because the Box-Cox method (Box & Cox, 1964; Venables & Ripley, 2002) suggested the logarithm as the most appropriate transformation. We excluded all response times smaller than 300 (an implausibly small value for question-response times), resulting in 5 excluded data points (i.e., 0.1% of the data). The distribution of residuals was approximately normal.

3.1.2 Results

Table 3.1: Question-response times from Swets et al. (2008), standard errors in brackets.

	ambiguous	high attachment	low attachment
superficial questions	1898 (101)	1826 (81)	1846 (88)
RC questions	2954 (185)	2801 (179)	2490 (137)

Table 3.2: Linear mixed-effects models coefficients, their SEs, and corresponding t-values, for the analysis of question-response times in the Swets et al. experiment.

	Est. (SE)	t
RC-questions	0.34 (0.05)	6.29
High-Low	0.01 (0.03)	0.3
Ambiguous-High	0.02 (0.03)	0.99
RC-questions × High-Low	0.07 (0.04)	1.98
RC-questions × Ambiguous-High	0.04 (0.04)	1.22

Table 3.1 shows the mean question-response times in the superficial and RC questions conditions. Table 3.2 shows the details of our linear mixed-effect model fit: we found significantly higher answering times in RC question conditions than in superficial questions conditions ($\hat{\beta} = 0.34$, $SE = 0.05$, $t = 6.29$). There were no significant differences between attachment conditions when questions were super-

ficial ($t = 0.3$, and $t = 0.99$). We found a significantly larger difference between high and low attachment conditions when questions were about the RC ($\hat{\beta} = 0.07$, $SE = 0.04$, $t = 1.98$), but no such effect for the difference between ambiguous and high attachment conditions ($\hat{\beta} = 0.04$, $SE = 0.04$, $t = 1.22$).

3.1.3 Discussion

The results of our analysis suggest that, while the difference between question answering times in high and low attachment sentences is higher for RC questions than for superficial questions, there is no evidence of an effect of question type on the difference between ambiguous and high attachment sentences. Although SDCF found significantly higher question response times in ambiguous sentences than in unambiguous sentences when the questions were about the RC, the lack of significant interaction in our analysis makes their finding difficult to interpret.

One possibility is that there is no difference between question answering times for ambiguous and high attachment sentences when questions are superficial, but that there is such a difference when questions are about the relative clause. If this were correct, the failure to find a significant interaction would constitute a type-II error. Another possibility is that the mental load associated with processing in the ambiguous condition slows down processing during the question-answering phase, irrespective of the kind of question that was asked. In that case, the failure to find a significant slowdown in superficial questions concerning ambiguous sentences relative to those concerning unambiguous sentences would constitute a type-II error. Both possibilities are compatible with the present results, and while the first option does not appear easily compatible with the URM, the second one does.

While the evidence is insufficient to decide between the two options, the second explanation is preferable on the grounds of parsimony, because SDCF's explanation of the question-response slowdown is based on the assumption that readers *may sometimes* resort to underspecification in spite of RC questions — a suboptimal strategy, when RC questions are asked. The mechanism leading to this behavior adds degrees of freedom to the underspecification model, which do not appear independently motivated. It is furthermore not easily compatible with the lack of an ambiguity advantage in the RC questions condition.

Next, we will discuss Swets et al.'s second finding: the lack of an ambiguity advantage in the RC questions conditions and its implications for the URM.

3.2 A Reinterpretation of the URM

3.2.1 Incrementality in Sentence Processing

Recall that van Gompel et al. (2000) argue that the ambiguity advantage occurs due to reanalysis cost in the disambiguated conditions. Interestingly, the ambiguity advantage effects in Traxler et al. (1998); van Gompel et al. (2001, 2005), as well as Swets et al. (2008) were not found during early reading of the disambiguating word. Instead, they were found either at the post-critical region, or in late reading time measures on the critical region. The URM can explain such late effects: first, the parser non-deterministically attaches the RC as soon as possible. Later, if the disambiguating word is incompatible with the initial analysis, it may trigger reanalysis, causing slowed reading of the disambiguating word, or in the post-critical region (due to spill-over). Thus, the URM offers a *strictly incremental* account of the ambiguity advantage: the parser does not postpone RC attachment, but carries it out as soon as it can.

The strategic underspecification account, in contrast, is not easily compatible with strictly incremental processing: Because disambiguation of a local ambiguity may occur on any word in the sentence, the good-enough parser must postpone the decision about whether to carry out RC attachment or to underspecify until enough evidence has been accumulated. In Swets and colleagues' sentences in (12) (repeated as 13), this happens at the reflexive *himself/herself*. Because the parser is assumed to underspecify only globally ambiguous sentences, the reflexive is the earliest point at which it can be sure that a sentence is globally ambiguous. Thus, Swets et al. (2008), must assume a certain amount of processing delay in order to explain that participants make use of different processing strategies in different attachment conditions. SDCF do not discuss this issue, and therefore it is not clear precisely how much delay they assume.

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

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

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

This necessary delay in RC attachment appears to be at odds with the incrementality

assumption made by most current theories of sentence processing with the notable exception of Construal (Frazier & Clifton, 1996, 1997). Typically, the parser is assumed to integrate every word into the current sentence structure as soon as it becomes available. For example, van Gompel and Pickering (2006) argue in favor of incrementality on the basis of the finding that sentences such as (14) lead to processing difficulty. This is arguably so because the string *The evidence examined* is initially interpreted as a main clause, while *examined* turns out to be the verb of a reduced relative clause later. Thus, the parser is forced to revise its initial erroneous decision when it encounters the disambiguating phrase *by the lawyer*.

Because the initial decision must have been taken before the disambiguating material is processed, van Gompel and Pickering (2006) argue that this finding constitutes evidence in favor of incrementality. The same kind of argument can be made on the basis of other garden-path effects as well.

(14) The evidence examined by the lawyer turned out to be unreliable.

However, the existence of garden-path effects does not provide evidence for *strict incrementality*, i.e., the assumption that every word and every phrase is immediately integrated into the structure of the current sentence to the fullest extent possible. It is possible that the attachment of relative clauses (or possibly of adjuncts in general; e.g., Frazier & Clifton, 1997) can be delayed to some extent. For example, attachment of a RC may be delayed until the RC has been fully processed.[2] The assumption of such delayed RC attachment is compatible with the findings concerning the ambiguity advantage (Traxler et al., 1998; van Gompel et al., 2001, 2005; Swets et al., 2008) because in these experiments, the ambiguity advantage effects were found either on the post-critical region (which coincided with the end of the relative clause), or in late reading times measures, such as total fixation time, which are likely to have been caused by regressions originating from the post-disambiguating or later regions.

[2] A possible alternative assumption is that RC attachment could be delayed until the RC verb, and possibly its core arguments, have been processed. Processing of the verb can be considered to be a minimal requirement for establishing a grammatical dependency between a relative clause and its attachment site. This is because in sentences like (13) the parser cannot establish grammatical number agreement between the head noun (*maid/princess/son*) and the relative clause verb, or assign a thematic role (e.g., agent or patient) to it the prior to encountering the verb. However, the question about the exact amount of delay is beyond the scope of the present thesis. In the following, *delayed RC attachment* will be taken to mean that the parser attaches the RC only once it has been fully processed.

3.2.2 URM without Reanalysis

Importantly, it is sufficient to assume that the parser cannot combine syntactically incomplete constituents in order to reconcile the URM with Swets et al.'s findings. A consequence of this assumption is that the parser needs to wait until the RC has been fully processed before attaching it. When the end of the relative clause is reached in ambiguous sentences, one of two possible attachments needs to be made. According to the URM, the parser tries to construct both attachments simultaneously and then terminates all structure-building as soon as the first one is constructed. Thus, the time to complete an attachment on a particular trial is equal to the completion time of the attachment process that is the fastest on this trial. Because the completion times of each process are assumed to vary from trial to trial, some attachment completion times come from one attachment process, and some from the other. It follows that the average time to complete an attachment is the mean of the shorter attachment times from all trials. By contrast, only one attachment can be made in an unambiguous sentence, so the average attachment time is the average time required to complete the relevant attachment process (high or low attachment).

Since the parser thus has a higher chance of completing attachment relatively early in the ambiguous sentences in (7c) than in unambiguous sentences in (7a,b), the average reading time in (7c) should be shorter than in (7a,b). Hence, no reanalysis cost needs to be invoked in the unambiguous conditions. Under this version of the URM, the locus of the ambiguity advantage is also the last word of the relative clause, which is *moustache* in the case of the sentences in (7). In order to distinguish this explanation of the ambiguity advantage from the reanalysis-based explanation, we will follow the terminology of Raab (1962) (see Miller, 1982, where this work is cited) by referring to this type of processing facilitation as *statistical facilitation*.

This version of the URM is compatible with Traxler et al.'s (1998) findings, as well as with the evidence for the ambiguity advantage found by van Gompel et al. (2000, 2001, 2005). Indeed, the reanalysis assumption in the original URM model is not only unnecessary but also renders the model's predictions impossible to quantify; in order to derive the predictions of a URM-with-reanalysis model, we would need to obtain estimates of the pure cost of reanalysis. Since the URM assumes non-determinism, the additional processing cost due to reanalysis in the disambiguated conditions cannot be used as an estimate of reanalysis cost. Without this information, it is impossible to derive quantitative predictions from the URM. Therefore, the statistical facilitation version of the URM is the only realization of the URM from which quantitative predictions can be derived, as we will do below.

While the strategic underspecification model does explain the effect of question type on the occurrence of the ambiguity advantage, SDCF's claim concerning the URM's predictions does not take an important fact about the URM into consideration. Although the URM always predicts a race in case of an ambiguity, it does not necessarily predict an ambiguity advantage of the same magnitude in all possible situations. A key observation about the URM's predictions (not discussed in the URM literature as far as we are aware) is that the predicted *amount* of statistical facilitation depends on the difference between the mean reading times of the two processes involved in the race. This is so because the degree of overlap between the completion time distributions of the structure-building processes engaged in the race depends on the difference between their mean completion times. The upper panel of figure 3.1 illustrates that a large amount of overlap between the two distributions (a small difference in means) leads to a high probability that one of the processes will finish relatively early. The lower panel of figure 3.1 on the other hand, illustrates that a small amount of overlap (a large difference in means) leads to a smaller probability of finishing early. When the overlap between the two distributions is small, the completion time distribution of the race process is largely identical to the distribution of the faster racing process.

Thus, the statistical facilitation predicted by the URM is largest when the difference between the mean completion times (MCT) of the racing processes is small, leading to a large overlap between the completion time distributions of the racing processes. In other words: the predicted ambiguity advantage is large when the two racing processes are equally fast or nearly equally fast and if there is a lot of variability in their completion times; when there is a large difference in between the MCTs of the processes, we may see only a very small or no ambiguity advantage at all.

Figure 3.2 illustrates how the MCT of a race between two processes depends on the difference between the MCTs of the respective racing processes. To illustrate this, we simulated a race between two processes with stochastically independent completion times.

While the MCT of one process was set to 600 ms, we increased the MCT of the other racing process from 600 to 900 ms, in steps of 50 ms. Based on the finding that the standard deviation of a reaction time sample appears to be approximately proportional to its mean (Wagenmakers & Brown, 2007), we set the SD of every simulated process to 10% of its MCT. Figure 3.2 demonstrates that the statistical facilitation is largest when the MCTs of the racing processes are equal, i.e., when the overlap between their completion time distributions is biggest. While there is

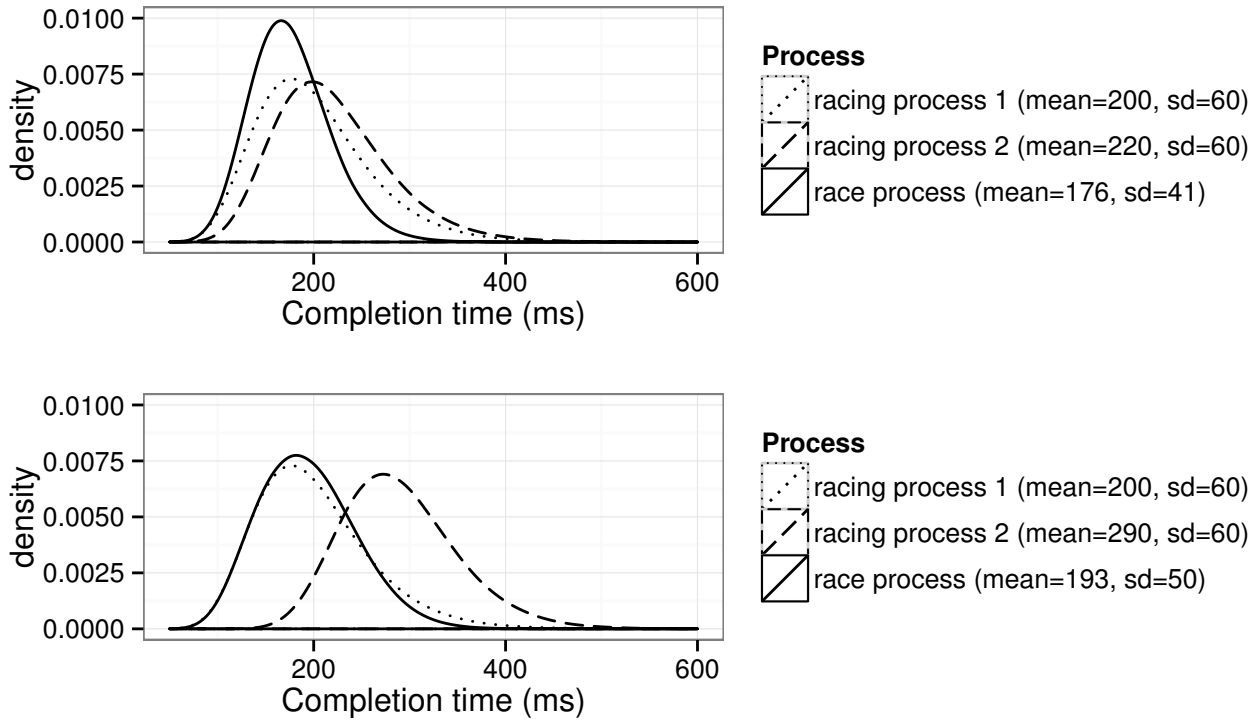


Figure 3.1: Simulated completion time distributions of racing processes and the resulting race completion times: a race process has lower mean completion times if there is a large overlap between the distributions of the racing processes (upper panel); if the overlap is small (lower panel), there is little to no facilitation. (The race process was simulated by repeatedly sampling one RT from each of the racing processes' completion time distributions, and using the smaller of the two numbers as the completion time of the race process. Reading times of both racing processes were assumed to be log-normally distributed (e.g., Ulrich & Miller, 1993; Limpert et al., 2001), with means and standard deviations as provided in the legend.)

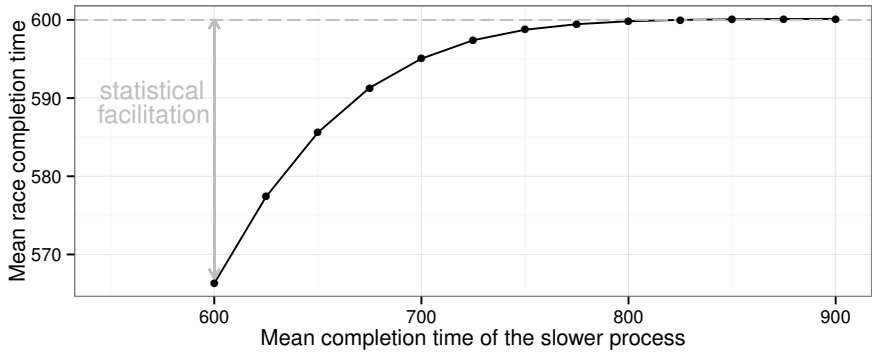


Figure 3.2: Mean completion time of a simulated race process as a function of the mean completion time of the slower of the two racing processes, while the mean completion time of the faster process remains at 600 ms. The mean race completion times are lowest when the difference between the completion times of the racing processes is small. (Simulations are based on a million samples drawn from log-normal distributions.)

a statistical facilitation of more than 30 ms when both racing processes are equally fast, its magnitude decreases as one of the racing processes becomes slower and the overlap between the completion time distributions of the racing processes decreases. For instance, when the slower of the two processes has a mean completion time of 750 ms, the statistical facilitation is reduced to less than 2 ms.

3.2.3 An alternative Explanation of Swets et al.'s Reading Time Findings

This relationship between the magnitude of the statistical facilitation and the difference between the mean completion times of the racing processes has direct implications for the URM predictions concerning the magnitude of the ambiguity advantage observed by Traxler et al. (1998): the statistical facilitation account of the ambiguity advantage predicts that the ambiguity advantage should decrease with an increasing difference between the mean completion times of the attachment processes.

Table 3.3 shows the reading times on the post-disambiguating region from SDCF. The reading times display an ambiguity advantage in the occasional and superficial questions conditions (i.e., the difference between low attachment and ambiguous conditions reading times: 36 ms and 52 ms respectively), and no ambiguity advantage in the RC questions condition (numerically, there was an ambiguity *disadvantage*

Table 3.3: Reading times on the post-disambiguating region, in ms (from the raw data of Swets et al., 2008). Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

	high attachment	low attachment	ambiguous
occasional questions	1203 (30)	1182 (50)	1146 (30)
superficial questions	1152 (27)	1118 (24)	1066 (23)
RC questions	1458 (44)	1233 (30)	1298 (40)

of 65 ms, which was not statistically significant). Table 3.3 also shows that, when questions concerned relative clause attachment, the post-disambiguating region was read more slowly in all attachment conditions. This is presumably because participants read the relative clause more carefully when they expected questions about it. Moreover, it appears that more careful reading increased the difference between reading times in high- and low-attachment sentences. High attachment sentences were read more slowly than low attachment sentences by 244 ms when RC questions were asked, while the reading times for such sentences differed by less than 30 ms when the questions were superficial or occasional.

Research on the effect of instructions on reading speed is consistent with the slowdown in unambiguous conditions. McConkie, Rayner, and Wilson (1973) found that participants read text passages faster if they were followed by superficial questions than if they were followed by questions requiring deeper semantic processing. In an eye-tracking experiment, Kaakinen, Hyönä, and Keenan (2002) found that readers spent more time reading a sentence when its topic was relevant to the task. In the light of these findings, slowed reading in the RC questions condition does not seem surprising. Since readers considered the relative clause more relevant when asked about its attachment after every sentence, they read the RC more carefully than when such questions were not asked. The increased difference in reading times between low and high attachment conditions is also not without precedent: Wotschack (2009) found eye tracking evidence for an interaction between question difficulty and the effect word predictability. When questions were more difficult, the effect of word predictability was more pronounced. It appears that in Swets et al.'s experiment too, the magnitude of difference between reading times in low and high attachment conditions depended on the depth of processing.

The important point is that the size of the difference in mean completion times of the two racing processes has direct implications for the magnitude of the ambiguity

advantage predicted by the URM. Because the race process depends on the completion time difference between low- and high-attachment, the URM actually predicts a pattern very similar to SDCF's core finding: it predicts a very small ambiguity advantage in the RC questions condition. That is because the difference in reading times between high- and low-attachment conditions can be considered an estimate of the difference between the completion times of the attachment processes engaged in the race. Since this difference is larger in the RC questions condition, the URM predicts a much smaller ambiguity advantage, as illustrated in figure 3.2.

3.3 Model 1: URM and the Effect of Task-demands on Reading Times

3.3.1 Method

In order to establish whether SDCF's results are in principle compatible with the URM, we used the mean reading times at the post-disambiguation region in the unambiguous conditions to predict the mean reading times in ambiguous conditions. While the post-disambiguating region consisted of between one and five words (the average length was 2.8 words)[3], we assumed for our simulation that the disambiguation took place on only one of these words. Therefore, we estimated the attachment-unrelated reading time in each question condition as 1.8/2.8 times the mean reading time in the (preferred) low attachment sentences and assumed that the standard deviation (SD) of the attachment process is 25% of the mean reading time in that condition, based on Wagenmakers and Brown's (2007) finding of a linear relationship between mean and SD of reaction times. We chose to set the SD to an arguably plausible value of 25%, as it fit the SDCF's results well and thus allowed us to illustrate that the URM is in principle compatible with SDCF's result.[4]

[3] Many thanks to Benjamin Swets for kindly providing us with the stimuli and data of the original study.

[4] We could not use SDCF's raw data to estimate the within-participant variability in attachment completion times due to the fact that their items were not matched for length of the post-disambiguating region or for the length and frequency of the words therein. The resulting between-items variability would have lead to overestimates of the within-participant variability in the completion times the attachment process. Furthermore, this problem is exacerbated by Wotschack's (2009) finding of an interaction between lexical variables and task difficulty. This effect would selectively increase estimates of variability in the RC question conditions and thus lead to underestimates of the predicted reading time in the ambiguous condition.

Table 3.4: Reading times for ambiguous sentences from the SDCF experiment and predictions of the URM (in ms), standard errors in brackets.

	ambiguous	ambiguity advantage	predicted ambiguous	predicted ambiguity advantage
occasional questions	1146 (30)	36	1132 (3)	50
superficial questions	1066 (23)	52	1076 (3)	42
RC questions	1298 (40)	-65	1222 (4)	11

3.3.2 Results

We simulated the predictions of the URM for ambiguous sentences in each question condition by repeatedly sampling pairs of values from two log-normal distributions (e.g., Ulrich & Miller, 1993; Limpert et al., 2001) corresponding to the low and high attachment conditions and using the smaller of the two. We sampled 576 such values (corresponding to 48 subjects with 12 sentences per condition) one million times. The reading times for the ambiguous condition from SDCF's experiment, as well as the predictions of the fitted race model are provided in Table 3.4. According to our simulation, the race model predicts an ambiguity advantage between 40 and 50 ms when questions are occasional or superficial, but a much smaller ambiguity advantage of 11 ms in the RC questions condition. An effect of this magnitude was unlikely to be detected in SDCF's experiment, and while the predicted magnitude of the ambiguity advantage differs in sign from the one obtained in the experiment, the predicted mean RT of $1222ms$ falls within the 95% confidence interval of the RT in the ambiguous condition (1220; $1376ms$). As such, the pattern of results found by SDCF can, in principle, be explained by the race model. This makes their argument against the URM much less compelling.

3.3.3 Discussion

SDCF argue that the parser's behavior at choice points is task-dependent. The parser underspecifies ambiguities unless the task requires ambiguity resolution, in which case the preferred reading is chosen. Indeed, the parser's behavior *does* seem task-dependent; this is clear from the effect of question type on overall reading times in unambiguous sentences. Harder questions do indeed seem to result in 'deeper' processing. However, this does not mean that its treatment of choice points is task-dependent. The present simulation shows that the lack of an ambiguity advantage

in the RC questions condition could simply be a consequence of deeper processing, which appears to increase the difference between the completion times of high- and low-attachment processes. Therefore, SDCF's evidence does not necessarily entail that the parser's *treatment of ambiguities* directly depends on task demands. Thus, task demands might be modulating the mean completion times in the manner discussed above, rather than modulating the parser actions per se; if so, the URM can explain the SDCF results.

In sum, the SDCF findings cannot distinguish between the task-demand explanation and the unrestricted race model account based on reading times alone.

3.4 Summary

In this chapter, evidence against the URM was inspected in more detail. It was shown that Swets and colleagues' results are compatible with a version of the URM which assumes delayed RC attachment. Specifically, it was shown that there is an alternative explanation for what SDCF consider evidence for the influence of task demands upon the strategy employed for ambiguity resolution. Thus, with a minor modification, the URM can explain SDCF's result on the basis of the fact that the magnitude of the ambiguity advantage depends on the mean completion times of the processes underlying the race. The larger the difference between the completion times of the high and low attachment processes, the smaller the ambiguity advantage predicted by the URM, everything else being equal. Henceforth, the term URM will be used synonymously with this modified model. Furthermore, SDCF's evidence from question-response times was re-examined. It was shown that their finding is inconclusive, and is compatible with alternative explanations.

In sum, while SDCF's findings pose a challenge to the original formulation of the URM, they are compatible with the modified version presented in this chapter. The results of the reanalysis of question-response latencies remain compatible with the strategic underspecification proposal. Because the original formulation of the underspecification model is somewhat vague, the next chapter will attempt to sharpen the notion of underspecification by formulating two precise versions of what exactly the underspecification model might do. An attempt will be made to determine which of the two versions is in closer agreement with Swets et al.'s data, and to also determine whether underspecification is deterministic or non-deterministic.

Chapter 4

What is Underspecification?

The underspecification model proposed by (Swets et al., 2008) assumes that readers underspecify the RC attachment in the superficial questions condition because the task does not require them to carry out RC attachment. The result is an ambiguity advantage: ambiguous sentences are read faster than their unambiguous counterparts. In the RC questions condition, no ambiguity advantage occurs because readers underspecify very rarely. However, because they do underspecify sometimes, RC questions about ambiguous sentences are slower than about their unambiguous counterparts.

Importantly, SDCF's underspecification account also predicts a small ambiguity advantage in the RC questions condition. This is because it assumes that in RC questions conditions, readers assign ambiguous sentences the preferred structure on most occasions, but that on other occasions, they underspecify attachment in spite of the task. As a result, the average reading time in ambiguous conditions should be shorter than in the preferred condition, because readers do underspecify on some trials. The assumption that readers sometimes underspecify in spite of the task is necessary in order to explain the slower answering of RC questions about ambiguous sentences compared to unambiguous sentences. Therefore, from the perspective of both models — URM as well as strategic underspecification — the failure to find an ambiguity advantage during reading in the RC questions condition must be considered a statistical type-I error. However, while the URM predicts a relatively small ambiguity advantage in the RC questions condition, which is likely to go undetected, the underspecification model predicts it to be of a somewhat larger magnitude: recall that the explanation of the slow-down in question answering-times rests on the assumption that there are some trials on which the RC attachment

operation is omitted during reading, and later carried out during the question-answering phase of the trial. Under these assumptions, the additional time required to answer RC questions about ambiguous sentences (compared to their unambiguous counterparts) provides us with an estimate of the speed-up predicted during reading.

Assuming that RC attachment requires the same amount of time, irrespective of when it takes place (i.e., during reading or during question-answering), the difference between question-response times for ambiguous and unambiguous sentences in table 3.1 on page 16, suggests that the magnitude of the expected ambiguity advantage is at least $150\,ms$ (RT for ambiguous minus RT for high attachment sentences). The failure to find such a large effect appears relatively unlikely.

However, underspecification does not have to predict such a straightforward relationship between reading times and question-answering times. In other words: alternative models of underspecification are possible. In the following, two such models will be presented, one of which is a formalization of Swets et al.'s model, while the other is a more parsimonious modification of this model.

The aim of this chapter is to clarify the precise assumptions underlying underspecification, and to attempt to distinguish between them empirically, using Swets et al.'s data. In other words, it will be attempted to answer the following questions: (a) what exactly happens during underspecification trials; (b) how rarely does underspecification occur; (c) do most participants employ underspecification on some occasions, or do some participants make use of underspecification all the time, whereas others never do?

Before the alternative models of underspecification are presented, it is important to understand the salient facts of the Swets et al. study first, which will be used to motivate certain assumptions of the models. As such, the models are both based on *post-hoc* assumptions, but may provide us with valuable insight given those assumptions.

4.1 Overview of the Relevant Findings

Here, a summary of some of the relevant findings concerning response accuracy, question-answering time, and reading time data from the RC questions condition in Swets et al.'s experiment will be presented.[1] Only the data from the RC questions condition was analyzed because on each trial in this condition three dependent

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

measures were recorded which are pertinent to underspecification: reading time, question answering time, and RC attachment indicated by the response (high attachment, or low attachment). Because across all question conditions (ocasional, superfical and RC questions), Swets et al. (2008) found effects of attachment on the potentially disambiguating word (*himself/herself*) and the spill-over region (*in public*), the time participants required to read both regions (treated as one region) was analyzed as a reading time measure. In the analysis, all trials with question answering times of less than 15 seconds were used.[2]

The data of 11 out of 48 participants were excluded prior to analysis, 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 high attachment condition, and 6 due to errors in the low attachment condition. Their data was excluded because such high error percentages may be indicative of a reading strategy in which readers consistently attach either high or low, 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 4.1 shows the average reading time at the critical region, *himself/herself in public*. It shows that the high attachment condition is read more slowly than the ambiguous and the low attachment conditions. This could be either because (a) the parser always attempts to construct a low attachment reading first, even in the high attachment conditions; or (b) the first noun requires more time to be retrieved from memory than the second noun, because the former is more distant from the relative clause. Although ambiguous sentences are read somewhat more slowly than low attachment sentences, the difference is not significant (the 95% confidence interval for reading times in the ambiguous conditions is $[1832\,ms; 2084\,ms]$). Table 4.2 shows the average question response time by attachment condition. Question-responses in ambiguous conditions are slower than in unambiguous conditions, and questions about low attachment sentences are answered faster than questions about high attachment sentences.

Table 4.1: Mean reading times (in milliseconds) for the critical region, by attachment. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

high attachment	low attachment	ambiguous
2143 (63)	1845 (44)	1958 (63)

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

Table 4.2: Mean question answering times for RC questions, by attachment. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

high attachment	low attachment	ambiguous
2826 (98)	2512 (86)	3033 (116)

Table 4.3 shows the average proportions responses indicating low attachment by attachment condition. For example, a 'yes'-response to a high question is considered to indicate high attachment, while a 'no'-response is considered to indicate low attachment.

While participants answered questions about unambiguous sentences with an accuracy of approximately 80%, the percentage of responses indicating low attachment in ambiguous sentences was closer to 50%, suggesting that the preference for low attachment was relatively weak.

Table 4.3: Mean proportions of responses indicating low attachment by attachment condition. Standard errors in brackets.

high attachment	low attachment	ambiguous
0.22 (0.02)	0.83 (0.02)	0.59 (0.02)

Table 4.4 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 high attachment sentences than those associated with correct responses. For low attachment sentences, the pattern is reversed. However, neither difference is statistically significant.

Table 4.4: Mean reading times in the unambiguous condition at the critical region by correctness of the response. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

	high attachment	low attachment
correct response	2165 (70)	1834 (47)
incorrect response	2064 (94)	1902 (81)

Table 4.5 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 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 4.5: Mean question-answering times in unambiguous conditions by attachment and correctness of the response. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

	high attachment	low attachment
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, high attachment conditions are read more slowly than low attachment and ambiguous conditions.

2. Question-response times in ambiguous conditions are slower than in unambiguous conditions, and questions about low attachment sentences are answered faster than questions about high attachment sentences.

3. In ambiguous sentences, the proportion of responses consistent with an low attachment was approximately 50%, suggesting a weak preference for low 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. In other words, reading time was not affected by whether or not the question on that trial was answered correctly.

5. In contrast, longer question-response times were seen for incorrect responses to unambiguous conditions, compared to response times for correct responses.

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

4.2 Two Ways to Underspecify

Swets et al. claim that readers sometimes engage in RC attachment during question-answering (*question-triggered RC attachment*, henceforth) 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 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 the ambiguous sentence (12c) from Swets et al.'s experiment (repeated as (15)) looks like the one shown in figure 4.1a. We will refer to this kind of underspecification as a *partial specification*, because partial information about RC attachment is stored by the parser.

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

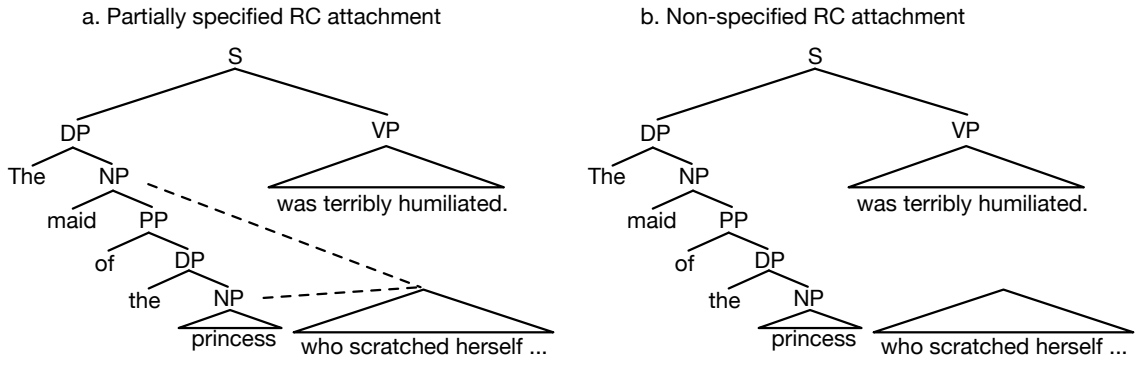
Importantly, the original ambiguity advantage found by Traxler et al. (1998) as well as the ambiguity advantage in the superficial questions conditions of the Swets et al. experiment is not straightforwardly compatible with partial specification. This is because the parser needs to store attachment-related information (engage in partial specification) in ambiguous as well as unambiguous conditions. Therefore, underspecification will be predicted to be faster than regular 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 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. Figure 4.1b illustrates the resulting structure of sentence (15). 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 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 question-triggered RC attachment is not possible, at least not without a prohibitively expensive repars-

[3] An alternative 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.

Figure 4.1: Underspecification and non-specification.



ing process. This is because the parser does not know what the available attachment sites are. Therefore, in trials where the comprehender engages in non-specification, they have to resort to guessing the answer to the questions.[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 4.5, 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-

[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*. Thus, a non-specified representation does not allow the reader to distinguish between grammatical and ungrammatical attachment sites for the RC.

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

specification models.

4.3 Two Models of Underspecification

4.3.1 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 4.2a shows the logic of the partial specification model required to account for Swets et al.'s results in the RC questions condition. The figure shows that when attachment is unambiguous, readers have only one option: attaching the relative clause. However, attaching the relative clause correctly should lead to only correct responses to questions; that is clearly not the case, since participants do give incorrect responses. There are two possible explanations for incorrect responses: one is that readers process the sentence in a much more shallow manner, and then try to guess the correct answer to the question. A second explanation is that, although readers always process unambiguous sentences in the same manner, they sometimes fail to retrieve the sentence representation during the question answering process. As a result, readers try to guess the correct answer. The first explanation predicts that incorrect responses should be preceded by faster reading, while the second explanation predicts no such difference. Although the pattern in table 4.4 is inconclusive in that respect because it provides no evidence of a significant speed-up on trials followed by incorrect responses, it is numerically closer to the predictions of the second explanation, i.e., the one assuming no shallow processing. We will therefore assume that, even on trials where an incorrect response is given, readers process the sentence by fully specifying the attachment, but end up failing to retrieve the fully specified structure and have to resort to guessing during question-answering. Importantly, we will assume that guessing requires more time than regular question-answering, as discussed above in connection with the pattern in table 4.5.

By contrast, when attachment is ambiguous, participants underspecify on some trials by carrying out partially specified RC attachment. Therefore, they read fast but answer questions slowly, because they need to carry out RC attachment before responding. On non-underspecification trials, they carry out RC attachment during reading and therefore read more slowly, but are fast at answering questions. Irrespective of whether the sentence representation is fully or partially specified, its retrieval could fail during the question-answering phase; in such a case, a guess as

to the correct answer is generated.

Because we assume that guesses are the result of a failure to retrieve or to create an unambiguous sentence structure, we will also assume that RC attachment during the question-answering phase (*question-triggered RC attachment*) and guessing are mutually exclusive. This is because the failure to retrieve a partially specified sentence structure precludes comprehenders from arriving at a fully specified structure. This unavailability of a fully specified structure, in turn, results in a guess.

We will also assume, in agreement with the results in table 4.3, that the probability of failure does not depend on the attachment condition or the attachment-related operation carried out during reading (high attachment, low attachment, or underspecification). Importantly, we need to assume that the parser can choose to attach high or low in order to account for the fact that the preference for low attachment in ambiguous conditions is very weak (question-responses indicate low attachment on only 58% of the trials, according to table 4.3).

To summarize, the partial specification model makes the following assumptions about the parser's operations:

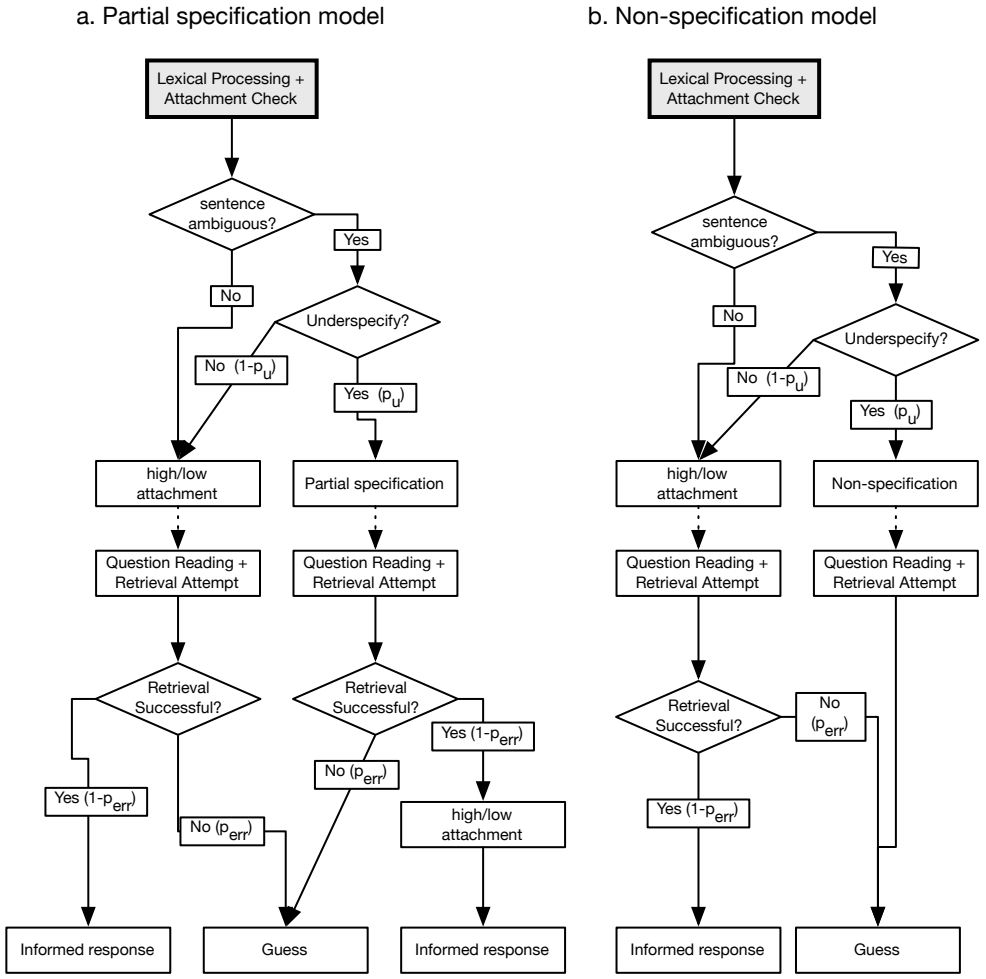
1. Readers always fully specify RC attachment in unambiguous sentences, but in ambiguous sentences they may choose to underspecify it with probability p_u. When readers do not underspecify the attachment, they can choose to attach the RC low (with probability p_{low}) or high (with probability $1 - p_{low}$).

2. Answering questions about RC attachment requires the retrieval of (parts of) the sentence representation. Retrieval of a sentence representation may fail with probability p_{err}, irrespective of whether it is fully or partially specified.

3. When retrieval fails, comprehenders attempt to guess the answer. Although they may have a bias towards 'yes' or 'no' responses, the equal proportions of high and low attachment questions in Swets et al.'s experiment ensure that the probability of responses compatible with high and low attachment respectively is equal.

4. 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 the underspecified representation is successfully retrieved.

Figure 4.2: A flow-chart of the trial structure according to the partial specification model (left panel), and according to the non-specification model (right panel). Probabilities of decisions in brackets where appropriate.



5a. When the parser underspecifies, it stores information about potential attachment sites, thus allowing for question-triggered RC attachment if necessary. We call this assumption 5a because the alternative proposal (presented below), will make a different assumption (5b).

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

1. Partial specification requires less time than full specification of high or low attachment.

2. Low attachment requires less time than high attachment. This assumption is motivated by the findings presented in table 4.1.

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 4.5.

4.3.2 Non-Specification

In the alternative non-specification model, we will assume 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 the appropriate attachment, as illustrated in figure 4.2b. In most cases, this attachment is followed by a correct response to the comprehension question, but in some cases, participants have to guess the answer due to a failed retrieval from memory.

During reading of 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 respond correctly 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. 4.1a), the non-specification model assumes that no such information is stored (as in fig. 4.1b). The consequence of this 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 some guesses and some informed responses preceded by RC attachment during the question answering phase.

This model can account for the findings presented so far. It can explain the higher question response times in the ambiguous condition because participants have to resort to guessing more often in that condition; by assumption, the guessing process consumes more time. Because it takes longer to generate a guess than to compare the sentence representation to the content of the question, this model predicts that there should be larger proportion of relatively long reaction times among questions

concerning ambiguous sentences compared to reaction times for questions concerning unambiguous sentences. Furthermore, the high proportion of trials involving guessing also explains the fact that there is no strong preference towards high or low attachment in ambiguous conditions.

The non-specification model makes the same assumptions about the timing of processes as the partial specification model, as well as assumptions 1-4 about the parser's operations. Instead of assumption 5a, however, the non-specification model adopts assumption 5b.

5b. When the parser underspecifies, it does *not* store any information about potential attachment sites, and thus does not allow for question-triggered RC attachment. As a result, the only way to answer a question on underspecification trials is to underspecify.

It should be added that, unlike Swets et al.'s original underspecification model, neither of these models predicts an ambiguity advantage of 150 ms or more. This is because both models assume that readers can assign ambiguous sentences high- or low-attachment structure. While Swets et al.'s model assumes that reading times at the critical region stem from a mixture of low attachment and underspecification trials, the present models assume that reading times at the critical region are composed of high attachment, low attachment and underspecifiction trials. Because high attachment requires more time than low attachment, the exact proportion of low attachment trials in ambiguous sentences will determine whether ambiguous sentences are read faster or slower than the preferred low attachment condition on average.

4.4 Models 2 and 3: Modeling Underspecification

The partial specification model and 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, the models make different predictions about the timing 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 low attachment trials should be equal to the difference in response times between low attachment trials and underspecification trials.[5] Furthermore, underspecification trials followed by question-triggered high or low attachment should result in responses indicating such attachment.

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 high and low attachment with equal probability.

Because these predictions cannot be tested without obtaining estimates of RC attachment duration, as well as of the proportion of underspecification trials and their response latencies, we formalized both models under the assumption that reading times and response times follow gamma distributions with one scale and different shape parameters, in order to obtain maximum-likelihood estimates of all model parameters and in order to compare the quantitative fits of the models to the data. The simultaneous estimation of model parameters and comparison of the best model fits can allow us to find out whether response times on underspecification trials are closer to the predictions of the partial specification model or to those of the non-specification model.

4.4.1 Method

We formalized both models as described in the appendix in order to obtain log-likelihood functions for each. We then fitted both models to each participant's data separately in order to account for between-participant variability in speed, attachment preferences and error rates. We assumed that all reading times and reaction times follow a gamma distribution with (1) a common scale parameter and separate shape parameters for (2) base reading time per word (i.e., underspecification), (3) high attachment, (4) low attachment, (5) informed question answering, and

[5] 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.

(6) guessing. Furthermore, we estimated (7) the probability of failing to recall a sentence during question-answering, (8) the probability of underspecifying in the ambiguous condition, (9) the probability of choosing an low attachment when attaching the RC in the ambiguous condition. In addition, we fit constrained versions of both models under the assumption that every participant pursues a particular strategy on all trials. In this additional set of models, we constrained the probability of underspecifying an ambiguous sentence to be either 1 or 0 for each participant. Consistent underspecification corresponds to a value of 1, while lack of underspecification corresponds to a value of 0.

Search for the maximum-likelihood estimates (e.g., Myung, 2003) of the parameters was carried out using the SUBPLEX algorithm (Rowan, 1990), which is a modification of Nelder and Mead's (1965) SIMPLEX. We conducted the optimization in GNU-R (R Core Team, 2013) using an interface to the *NLopt* package (Johnson, 2013). The log-likelihood of each trial was computed the sum of the log-likelihoods of all three dependent variables for each trial (i.e., reading time, question-answering latency, response), given the parameters.

We used the Bayesian Information Criterion (BIC) as well as the BIC approximation to the Bayes factor (Wagenmakers, 2007) for model comparison. The BIC is computed according to equation 4.1 and is a function of the maximized log-likelihood of a model ($log\ L$) and the number of free parameters (n_{par}), as well as the number of observations (n_{obs}). It increases with the number of free parameters, and decreases with increasing log-likelihood. Therefore, model quality in terms of the most parsimonious fit to the data is better for lower BICs. Because we modeled the dependency between three dependent variables (reading time, response time, and response), we set n_{obs} for each participant to the number of trials (36) times three. In addition, we computed the BIC approximation to the Bayes factor according to equation 4.2 (Wagenmakers, 2007), where BIC_1 and BIC_2 are the BIC values of the models to be compared. The Bayes factor (BF) quantifies the evidence in favor of model 1 over model 2. By convention, the evidence is considered *weak* when $1 < BF < 3$, *positive* when $3 < BF < 20$, *strong* when $20 < BF$ (Raftery, 1995).

$$BIC = -2 \cdot log\ L + log(n_{obs}) \cdot n_{par} \tag{4.1}$$

$$BF_{12} = e^{(BIC_2 - BIC_1)/2} \tag{4.2}$$

Table 4.6: Unconstrained models: log-likelihoods, average BIC, number of participants for whom the model provides the best fit with a BF of 3 or more, and number of free parameters per participant.

	model	$logL$	$\overline{BIC}$	$n_{selected}$	n_{par}
1	Non-specification	-22158	1240	2	9
2	Partial specification	-22164	1240	2	9

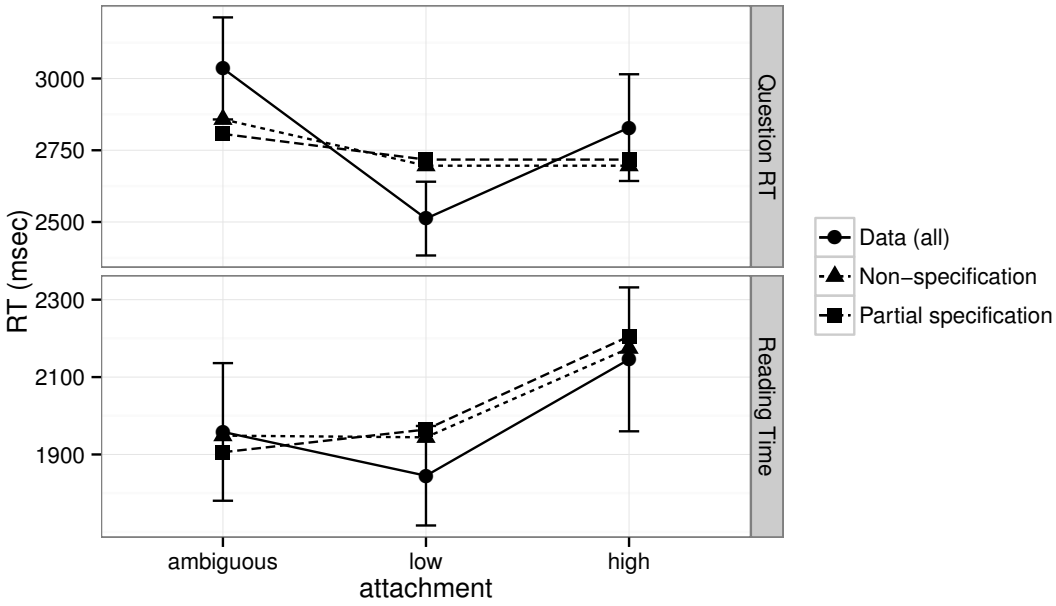
4.4.2 Results and Discussion

Table 4.6 shows the sum of the maximized log-likelihoods, the average BIC for that model, and for how many participants this model was considered to provide the best fit with a BF of 3 or more.[6] Table 4.6 shows that the log-likelihoods, as well as BICs of the two models are very close, with a slightly better fit for the partial specification model corresponding to higher log-likelihood. The fact that non-specification provided the best fit (with a Bayes factor of more than three) for only 2 out of 37 participants, while partial specification had the best fit for another 2 participants is in line with that. These results indicate that both models can fit the present data set equally well. In the light of these results, we cannot make inferences as to which model describes the data better.

Figure 4.3 shows the average predicted reading times and question latencies vis à vis the data. Figure 4.5 shows the percentages of responses indicating low attachment and the models' predictions thereof. The figures show that both models are able to capture the data to an equal extent. Both partial-specification and non-specification can account for a slowdown in question-response times, while accounting for the response patterns and the differences in reading times. However, both models appear to under-predict its magnitude. Further analyses revealed that the mean question-answering times in the ambiguous condition were in closer agreement with the data when all trials with reaction times above 12 *sec* (9 trials) were excluded from the data. The average predicted RTs are shown in figure 4.4. Thus, the numerical discrepancy appears to be driven by outliers, and may not arise if a more outlier-sensitive measure of deviance such as root-mean square deviation (RMSD) or (adjusted) R^2 were used.

[6] We disregard all preferences with a BF of less than three for two reasons. Firstly, even if overall, the two models were equally good fits for a given data set overall, some within-participant differences in log-likelihood for the two models can be expected due to sampling variability. Secondly, even if the actual maximized log-likelihoods for one participant's data were the same, minor log-likelihood differences between two models can be expected due to the fact that the numerical optimization algorithm (SUBPLEX) we used does not find the *exact* but rather an approximate maximized likelihood.

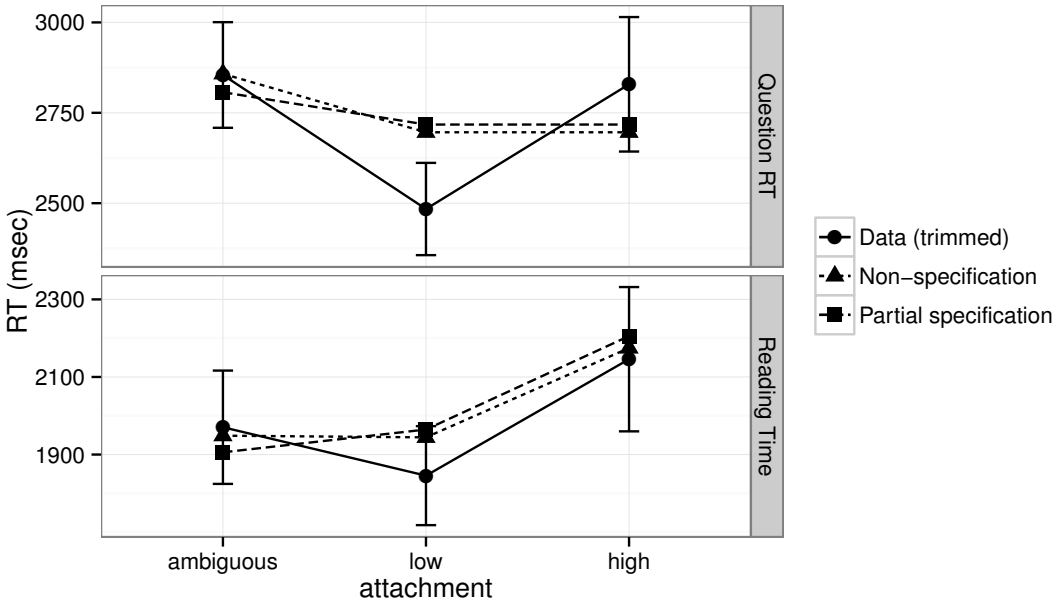
Figure 4.3: Model predictions in comparison to question-answering latencies (upper panel) and reading times (lower panel), based on all trials. Error bars correspond to standard errors.



We inspected the by-participant estimates of the probability of underspecification as well as the probability of choosing low over high attachment in an ambiguous condition when RC attachment is carried out. This was done in order to assess the possibility that every participant consistently used a particular strategy, e.g., always underspecified or never underspecified. The estimates are shown in figure 4.6, where each point corresponds to the estimates of these probabilities for one of the models. While the left panel of figure 4.6 suggests that more participants tended to choose low over high attachment, the lack of extreme values (zero or one) suggests that according to both models, most participants assigned ambiguous sentences different structures on different occasions. The estimates of the underspecification probability in the right panel of figure 4.6, however, are often either zero or one. Under the assumptions of the present models, this finding suggests that some participants never underspecified and some always underspecified.

According to the non-specification model 24 out of 37 participants never underspecified, as their estimate of the probability of underspecification was above 0.99. At the same time, only one participant always underspecified, as their estimate of the probability of underspecification was below 0.01. According to the partial specification

Figure 4.4: Model predictions in comparison to question-answering latencies (upper panel) and reading times (lower panel), based on trials with question-response RTs below 12 *sec*. Error bars correspond to standard errors.



model, 20 out of 37 participants never underspecified, while 11 always underspecified. In order to assess whether the data of participants without such extreme estimates provides evidence against the idea that they pursued such a strategy too, we fitted the non-specification and partial specification models to the data under the assumption that the probability of underspecification can be either one or zero.

Table 4.7: Constrained models with a discrete attachment parameter: Log-likelihoods, average BIC, number of participants for whom the model provides the best fit with a BF of 3 or more, and number of free parameters per participant.

	model	$logL$	$\overline{BIC}$	$n_{selected}$	n_{par}
1	Non-specification	-22165	1240	2	9
2	Partial specification	-22165	1240	3	9

Table 4.7 shows the log-likelihoods and BICs for both models under the assumption that each participant either consistently underspecified or consistently attached the RC during reading. The log-likelihood difference between these models with a constrained underspecification probability parameter and their unconstrained counterparts in table 4.6 appear to be negligible.[7] This finding suggests that, under

[7]Unfortunately, we cannot use the BIC to choose the more parsimonious model in this case,

Figure 4.5: Data and model predictions for the percentage of responses indicating low attachment. Error bars correspond to standard errors.

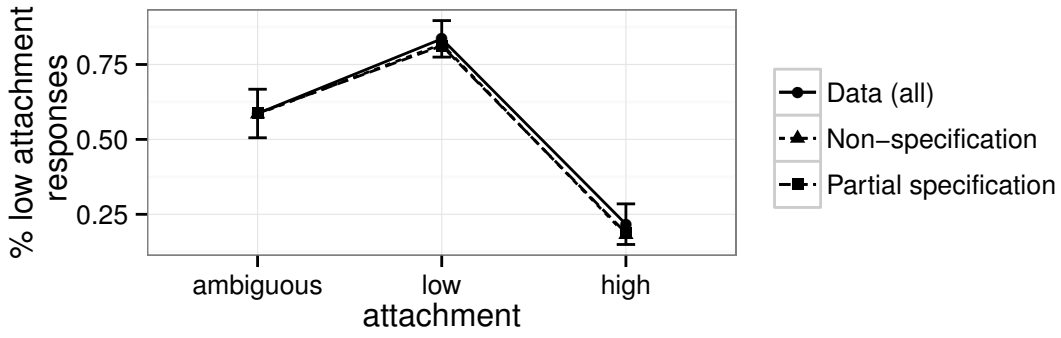
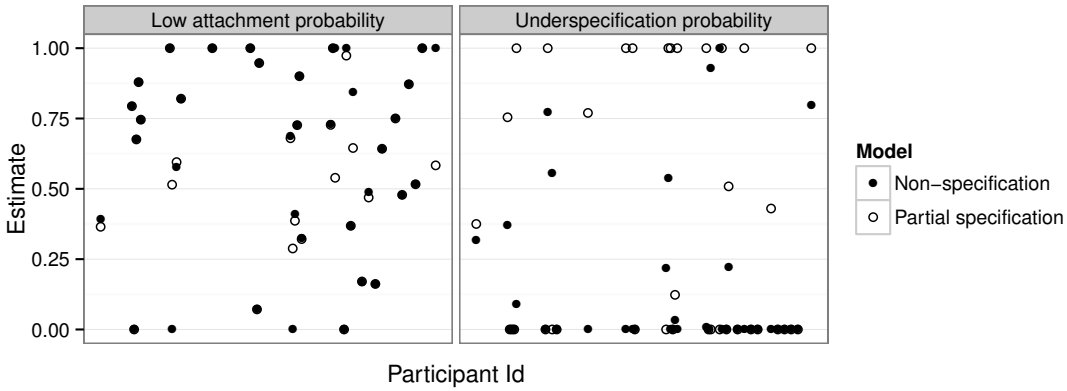


Figure 4.6: Scatterplot of estimates of underspecification probability and low attachment probability for both models.



the assumptions of both models, each reader likely pursues an underspecification or a non-underspecification strategy on all trials. According to the constrained non-specification model 33 out of 37 participants never underspecified and 4 always underspecified. According to the constrained partial specification model, 23 out of 37 participants never underspecified, while 14 always underspecified. In sum, both models seem to suggest the same conclusion.

because although the models in table 4.7 are more constrained, they have the same number of parameters as those in table 4.6. Thus, the BIC cannot take into account the fact that the constrained models are less flexible. Model selection techniques which assess model flexibility instead of the number of free parameters, such as Bayesian model selection or the Deviance Information Criterion (e.g., Spiegelhalter, Best, Carlin, & von der Linde, 2002) can be used to address this issue. However, this is beyond the scope of the present work.

4.5 General Discussion

We have presented two models which can account for 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 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 used maximum-likelihood estimation to compare the goodness of fit of the two models, and found that both models can account for the data equally well. Interestingly, we also found that the parameter estimates of both models suggest that some readers appear to consistently underspecify attachment, whereas others never underspecify. Thus, to the extent that either of our models provides an adequate description of the processes unfolding during relative clause attachment we can conclude that only some readers underspecify ambiguous structures. One possible reason is that while some participants attempt to finish the experiment as quickly as possible or with minimum mental effort, others attempt to fulfill the experimental task diligently. While the former make use of underspecification, possibly as part of larger strategy to save time, the latter fully specify attachment. However, the exact triggers for underspecification remain unclear as of yet.

Our failure to find a clear preference may be due to one or several of the following reasons. Firstly, the estimates of the model parameters suggested that the majority of participants did not underspecify. Thus, according to our estimates, the data of 18 participants or less was pertinent to the comparison between non-specification or underspecification. Furthermore, our estimates of underspecification-related parameters were based on only 12 trials in the ambiguous condition. The situation

was further complicated by the fact that the critical region consisted of between 2 and 6 words of different lengths. The resulting effect of variability unaccounted for by our models may have masked any differences in the goodness of fit between the two models. 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.

4.6 Summary

In this chapter, several problems with Swets et al.'s model of underspecification were pointed out. Two alternative models of underspecification were presented, and their predictions were tested on the data of Swets et al. (2008). Both models are consistent with the pattern of results found by Swets and colleagues. Although the model comparison proved inconclusive, the findings in the current chapter demonstrate that, to the extent that underspecification is an adequate account of the ambiguity advantage, some readers consistently underspecify ambiguous structures while others never underspecify. In other words, underspecification appears to be deterministic. Moreover, the notion of underspecification was theoretically sharpened, and may thus allow for more rigorous experimental testing in future research. However, in the next two chapters, evidence against underspecification and in favor of the URM will be presented.

Chapter 5

Evidence from Turkish

In the second chapter, two models of ambiguity resolution, the URM and strategic underspecification were presented. In the two subsequent chapters, their assumptions were discussed in detail, and their ability to account for the present data was evaluated. The assumptions of both theories were re-evaluated, and as a result, both models were revised in order to account for the findings of Swets et al. (2008). While the URM needs to assume that RC attachment occurs at the end of the relative clause, the strategic underspecification model needs to assume that ambiguous sentences can receive an high-attachment as well as a low-attachment interpretation. Because both modified models are compatible with the experimental findings presented so far (Traxler et al., 1998; van Gompel et al., 2001, 2005; Swets et al., 2008), we cannot distinguish between the two models. This is in part because the results concerning question-answering latencies are somewhat inconclusive, and in part because both models are compatible with the findings concerning reading time.

Importantly, the reason that the two theories make the same predictions for relative clause in English (and German) is that both theories assume that the attachment-related processes happen during the reading of the relative clause. Recall the ambiguous condition from Traxler et al.'s (1998) experiment in (7) (repeated as (16)).

(16) The son of the driver *that had the moustache* was pretty cool. (globally ambiguous)

In (16), the word in the sentence at which the underspecification parser decides whether to underspecify or not is *moustache*. This also happens to be the word at which the URM parser starts a race between two possible attachment options (high and low). Therefore, the URM and the strategic underspecification model make

the same qualitative prediction for sentences like 16: reading should speed at the end of the relative clause in ambiguous conditions. The fact that the critical word is the same for both theories is due to the fact that English relative clauses are *post-nominal*, i.e., they follow the nouns they modify. Therefore, readers encounter the relative clause only after they have read both noun phrases. For the modified URM-parser, this is the earliest point in time at which any attachment can be made, under the assumption that incomplete constituents cannot be attached. Because both attachment options become available simultaneously, a race begins. For the underspecification parser, this is the earliest point in time, at which the ambiguity of the sentence can be determined, because disambiguating input could have been provided anywhere in the relative clause.

In languages with pre-nominal relative clauses, such as Turkish, these points do not coincide. Consider the ambiguous Turkish sentence (17), which contains the relative clause (*each other hit*). This sentence can either mean that the football players hit each other, or that the fans of the football players hit each other. In other words, the relative clause can either attach *locally* (i.e., to the first noun *football players*), or *non-locally* (i.e., to the second noun *fans*).

(17) Dün akşam, **[birbirini döven]**$_{RC}$ futbolcu-lar-ın
Yesterday evening, each other hit *football player*-PL-GEN
hayran-lar-ı stadyumu hemen terk etti.
fan-PL-POSS *stadium immediately leave did.*
'The fans of the football players who hit each other left the stadium immediately, yesterday evening.'

The earliest point in this sentence at which an attachment can be made is the first noun (*football players*). This is because after reading this noun, the parser has a relative clause as well as a potential attachment site for it. Therefore, a dependency between *football players* and *hit each other* can be established. However, the parser does not yet know whether the sentence is ambiguous, and importantly, it has good reason to believe that it might be. This is because the genitive case suffix *-ın* (or *-nun*) signals that the noun *football players* is part of a complex noun phrase and that the parser can expect another potential attachment site at the next word.

If readers really underspecify RC attachment in ambiguous sentences in order to reduce processing effort, one would expect them to first process the entire complex noun phrase *the fans of the football players*, and then to decide whether to attach or not. Importantly, Swets et al. (2008) have suggested that the underspecification parser has the ability to postpone RC attachment until a question about

the sentence needs to be answered. It appears plausible that such a parser must have the ability to postpone RC attachment until the second noun. At the second noun, the underspecification account predicts that ambiguous sentences should be underspecified, given the appropriate task demands, while RC attachment should be carried out in their unambiguous counterparts. Therefore, given the right task demands, we expect to find an ambiguity advantage in Turkish. It should occur on the second noun (*fans*) because this is the earliest point in the sentence at which the underspecification parser can decide whether to underspecify or to attach.

The URM, on the other hand, predicts no ambiguity advantage. This is because it tries to attach the RC as soon as possible. The earliest point in the sentence at which this can happen is the first noun (*football players*). Therefore, the URM predicts RC attachment to happen on the first noun in the ambiguous and the local condition. Thus, there should be no difference in reading time between these two conditions. In the non-local condition, however, the parser should attempt to attach the RC at the first noun phrase and fail, possibly at the expense of a slowdown. At the second noun phrase, however, the RC is predicted to be attached successfully, resulting in longer reading times for the non-local condition because in sentences with local RC attachment as well as ambiguous sentences, attachment has already been completed at the previous word.

In the next section, we will present an experiment that was designed to test these predictions.

5.1 Experiment 1

In the present self-paced reading experiment, participants read Turkish sentences with ambiguous and unambiguous RC attachment, and answered occasional superficial questions about them. Questions were kept superficial in order to parallel the task demands employed by previous experiments which demonstrated an ambiguity advantage (Traxler et al., 1998; van Gompel et al., 2001, 2005; Swets et al., 2008). All experimental sentences in this experiment had the structure of sentence 18, in which the head noun of the relative clause (*football player/fan*) performs the function of the subject of the embedded verb (*hit*). The grammatical object of all relative clauses was the reciprocal pronoun *each other*. RC attachment was disambiguated by manipulating the grammatical number of the noun. In the local attachment condition in (18b), for example, the word *fan* is singular, and so the relative clause must attach to *football players* because only one person cannot 'hit

each other'. In the non-local attachment condition in (18c), the word *football player* is singular, and so the RC must attach to *fans*. In the ambiguous condition in (18a), both nouns are plural, and so the RC can attach to either noun.

(18) Dün akşam, ...
Yesterday evening,

a. GLOBALLY AMBIGUOUS (PLURAL-PLURAL)

[birbirini döven]$_{RC}$ futbolcu-lar-ın hayran-lar-ı
each other hit *football player*-PL-GEN *fan*-PL-POSS

b. LOCAL ATTACHMENT (PLURAL-SINGULAR)

[birbirini döven]$_{RC}$ futbolcu-lar-ın hayran-ı
each other hit *football player*-PL-GEN *fan*.SG-POSS

c. NON-LOCAL ATTACHMENT (SINGULAR-PLURAL)

[birbirini döven]$_{RC}$ futbolcu-nun hayran-lar-ı
each other hit *football player*.SG-GEN *fan*-PL-POSS

... stadyumu hemen terk etti.
stadium *immediately leave did.*

'The fans of the football players who hit each other left the stadium immediately, yesterday evening.'

Unfortunately, reading times for plural and singular versions of a word are not easily comparable, because the plural versions are always longer due to the addition of the plural suffix *-lar* (or in some cases *-ler*). In order to control for the effect of word length, three control conditions (in (19)) were used, in which the relative clause was replaced by an adverbial modifier (*of Fenerbahçe*, which is a well-known Turkish football team). These additional conditions will serve as a baseline, any differences from which must be considered effects of RC attachment.

(19) Dün akşam, ...
Yesterday evening,

a. CONTROL, GLOBALLY AMBIGUOUS (PLURAL-PLURAL)

[Fenerbahçeli] futbolcuların hayranları
of Fenerbahçe football player.PL.GEN *fan*.PL.POSS

b. CONTROL, LOCAL ATTACHMENT (PLURAL-SINGULAR)

[Fenerbahçeli] futbolcuların hayranı
of Fenerbahçe football player.PL.GEN *fan*.SG.POSS

c. CONTROL, NON-LOCAL ATTACHMENT (SINGULAR-PLURAL)

[Fenerbahçeli] futbolcunun hayranları
of Fenerbahçe football player.SG.GEN *fan*.PL.POSS

... stadyumu hemen terk etti.
stadium immediately leave did.

'The fans of the football players from Fenerbahçe left the stadium immediately, yesterday evening.'

The underspecification account predicts a speed-up in the ambiguous condition at the second noun, compared to the local attachment condition. This prediction translates into an interaction between modifier type and the grammatical number of the second noun. Reading should slow down in RC attachment conditions when the second noun is singular. The URM makes predictions concerning the same interaction. Its prediction about the lack of an ambiguity advantage in Turkish translates into a predicted lack of an interaction. In other words, there should not be any significant interaction between modifier type and the grammatical number of the second noun in the RC attachment conditions.

Importantly, because the URM predicts no significant attachment-related speed-up; in other words, a statistical null-result is expected under this account. In order to differentiate between possible outcomes compatible with the URM and strategical undespecification, respectively, we need to quantify the expected magnitude of an ambiguity advantage. Swets et al. (2008) is to the best of our knowledge the only study which has demonstrated an ambiguity advantage in self-paced reading. They found ambiguity advantage effects of 52 ms and 36 ms in the superficial and occasional questions conditions, though the latter was not significant. Three other studies have found evidence for this effect in eye tracking experiments: the original Traxler et al. (1998) article presented ambiguity-related speedups between 92 ms and 32 ms in total reading times at the disambiguating region in two eye tracking experiments. Both experiments involved relative clause attachment. In further two experiments concerning attachment of prepositional phrases, van Gompel et al. (2001) presented effects between 41 ms and 122 ms, mostly in total reading times, and one in regression-path duration. Furthermore, van Gompel et al. (2005) demonstrated speedups between 26 ms and 86 ms in total reading time and regression-path durations. This pattern suggests that most ambiguity advantage effects appear to concentrate somewhere around 50 ms. Thus, under the strategic underspecification account, we will expect an attachment-related speedup of approximately 50 ms. Under the unrestricted race model, we will expect no such effect, i.e., a speed-up of approximately 0 ms.

5.1.1 Method

Materials

Forty-two experimental sentence sets were constructed, which looked like (18) and (19). All relative clauses comprised between two and three words and always started with a reciprocal pronoun. Sentences were divided into different lists according to a latin-square design, such that every participant read exactly one sentence from each sentence set, and seven sentences from every condition. Experimental sentences were intermixed with 65 unrelated non-experimental filler sentences, with an example in (20). Overall, every participant read 107 sentences over the course of the experiment. Each list was randomized prior to presentation.

(20) Düşmanlar çok daha büyük bir ordu ile köylere tekrar saldırmış.
The enemy very more big a army with villages again attacked.
'The enemy attacked the villages again with a much larger army.'

Questions about the sentence were asked on one third of all trials (14 questions about experimental sentences, 21 questions about fillers). All questions about experimental sentences targeted the main clause, and not the RC attachment. For example, for the sentences in (19), the question had the form of (21).

(21) Hayranlar evlerini mi terketti?
Fans to home Q *leave?*
'Did the fans go home?'

Participants

Thirty-six students of Anadolu University in Eskişehir, Turkey participated in the experiment. All participants were native speakers of Turkish; their age range was 19-29 years. One experimental session took approximately 40 minutes to complete. Participants were paid 15 Turkish Lira for participation.

Procedure

The task was self-paced non-cumulative word-by-word reading. Presentation and recording was done with the Linger software package, version 2.94 by Doug Rohde. At the beginning of a trial, all words on the screen were masked by underscores. Participants pressed the space bar to reveal the next word. As the next word appeared,

the current one was masked by underscores again. The time between key-presses was recorded as the reading time for the word. Each participant read 5 practice sentences before the start of the experiment. Participants answered questions with 'yes' or 'no' by pressing the corresponding button on the keyboard.

5.1.2 Results

We used the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in R (R Core Team, 2013) to fit linear mixed-effects models (Pinheiro & Bates, 2000; Baayen, 2008; Gelman & Hill, 2007) to the reading time data. To determine the appropriate transformation, we used the Box-Cox method (Box & Cox, 1964; Venables & Ripley, 2002). For most positions, $1/\sqrt{RT}$ was suggested as the appropriate transformation (λ was close to -0.5). To avoid numerically very small coefficients, and to allow for a natural interpretation of coefficient direction (i.e., positive coefficients correspond to slowdowns), we transformed all reading times with $-10^4/\sqrt{RT}$ prior to analysis.

All models included fixed effects of grammatical number using helmert contrasts, in which (i) the condition with a singular first noun was compared with the two other conditions (effect of *N1 singular*), and (ii) the condition with a singular second noun was compared with the ambiguous condition (effect of *N2 singular*). Furthermore, all models included an effect of modifier type and its interaction with attachment. For modifier type, control conditions were coded as 0, and RC conditions were coded as 1, in order to allow for easy interpretation of the estimates of the interaction terms. With such coding, the estimate of the interaction between modifier type and the *N1 singular* contrast can be interpreted as the partial effect of N1 unavailability for RC attachment, while controlling for the effect of length. The estimate of the interaction between modifier type and N2 singular can be interpreted as the partial effect of N2 unavailability for RC attachment.

We included random intercepts for participants and items, as well as maximal by-participant and by-item random slopes. We did not include correlations between random effects, because some models failed to converge with those parameters, or produced pathological estimates of such correlations (i.e., 1 or -1). However, according to Barr et al. (2013) models without correlations between random effects do not significantly differ from maximal models in terms of controlling the Type-I and Type-II error rates. We removed all exceptionally small (< 150 ms) and all exceptionally large reading times (> 3000 ms) prior to analysis. This resulted in the removal of 0.46%, 0.73%, 0.46% of the data for the first noun, second noun,

Table 5.1: Mean reading times for the critical regions by condition. Within-participants standard-errors in brackets (Cousineau, 2005; Morey, 2008).

modifier type	attachment	pre-critical	N1	N2	spill-over
control	ambiguous	507 (16)	543 (20)	621 (20)	598 (21)
control	local attachment	505 (12)	561 (29)	650 (23)	590 (23)
control	non-local attachment	501 (12)	544 (20)	654 (33)	605 (30)
RC	ambiguous	535 (15)	575 (22)	654 (22)	593 (18)
RC	local attachment	527 (14)	561 (17)	644 (26)	612 (21)
RC	non-local attachment	534 (13)	548 (17)	725 (28)	608 (23)

Table 5.2: Linear mixed-effects models coefficients and the associated SEs and t-values for the analyses of reading times at the critical regions.

	noun 1		noun 2		spill-over	
	Est. (SE)	t	Est. (SE)	t	Est. (SE)	t
N1 singular	2.73 (2.05)	1.3	-1.82 (2.22)	-0.8	-0.37 (1.86)	-0.2
N2 singular	-1.14 (3.67)	-0.3	1.53 (3.65)	0.4	-3.14 (3.44)	-0.9
RC	10.33 (4.54)	2.3	7.18 (5.02)	1.4	5.65 (4.68)	1.2
RC × N1 singular	-3.97 (3.32)	-1.2	6.93 (3.17)	2.2	0.65 (2.82)	0.2
RC × N2 singular	0.59 (5.05)	0.1	-6.00 (4.72)	-1.3	4.02 (4.47)	0.9

and the spill-over region. In addition, two reading times at the second noun were removed, which corresponded to outlying residuals as established by means of the *qqPlot* function from the *MASS* package (Venables & Ripley, 2002). However, the results of both models, with and without outlier removal, were almost the same. No further outlier removal was necessary. In all models presented, $|t| > 2$ and $|z| > 2$ correspond to a significant effect at a significance level of .05. In addition, confidence intervals (CIs) based on profile likelihood are provided for the effects of interest.

Table 5.1 shows the mean reading times for the critical positions by condition. The results of the linear mixed-effects models showed a significant slowdown at the first noun in the RC attachment conditions as compared to the control conditions ($\hat{\beta} = 10.33$, $SE = 4.54$, $t = 2.3$). Furthermore, a significant slowdown due to the unavailability of the first noun for RC attachment was found at the second noun, which manifested as a *RC* × *N1 singular* interaction ($\hat{\beta} = 6.93$, $SE = 3.17$, $t = 2.2$, CI=[0.7; 13]). The unavailability of the second noun for RC attachment did not have an effect: the interaction *RC* × *N2 singular* was not significant ($\hat{\beta} = -6.00$, $SE = 4.72$, $t = -1.3$, CI=[−15; 3.3]). The confidence intervals for *RC* × *N1 singular* and *RC* × *N2 singular* translated to approximately [2 *ms*; 34 *ms*] and [−35 *ms*; 8 *ms*]

on the untransformed time scale.[1] No other effects were significant.

Because the critical finding was a null-effect, we computed a Bayesian credible interval for the effect of *RC* × *N2 singular*. We used Stan (Stan Development Team, 2014) to estimate the parameters of a Bayesian linear mixed effects model with full random-effects structure for participants (Barr et al., 2013), but not for items. We obtained a credible interval for the effect of *RC* × *N2 singular* by MCMC-sampling from the posterior distribution. The obtained credible interval was [37; 6] on the transformed scale, and corresponded to approximately $[-79\,ms;\ 15\,ms]$ on the original time scale. The posterior probability of *RC* × *N2 singular* being positive was 0.11.

5.1.3 Discussion

The analysis showed a significant slowdown at the first noun in the RC attachment conditions compared to the control conditions. This finding is not surprising considering that relative clauses are syntactically more complex than the adverbial modifiers in this experiment, which always consisted of one word.

The underspecification account predicted an interaction between modifier type and the N2 singular contrast. Because we expected a slowdown of approximately $50\,ms$ in the non-local condition relative to the ambiguous condition, we predicted an interaction the *RC* × *N2 singular* of approximately $50\,ms$, reflecting slower reading times for sentences with only one attachment option due to underspecification in ambiguous conditions. No such effect was found. Moreover, the confidence interval for this interaction ($[-35\,ms;\ 8\,ms]$) is not compatible with a predicted slowdown of this magnitude. More importantly, neither is the Bayesian credible interval of $[-75\,ms;\ 15\,ms]$, which was computed for this parameter. Our finding is clearly incompatible with an ambiguity advantage of the expected magnitude. However, this finding is compatible with the URM, because it predicted the lack of an ambiguity advantage in Turkish and thus the lack of a significant interaction between modifier type and the *N2 singular* contrast, which is what we found.

Furthermore, the interaction *RC* × *N1 singular* provided evidence for a RC attachment-

[1] This is because in the linear mixed effects model for the second noun, the intercept estimate was −439.70 on the transformed scale, and the estimated effect of modifier type was 7.18. Given these estimates and the fact that we transformed RTs according to $-10^4/\sqrt{RT}$ prior to analysis, the estimate of an effect affecting only the RC attachment conditions on the original time scale is approximately $(-10^4/(-439.7 + 7.18 + \hat{\beta}))^2 - (-10^4/(-439.7 + 7.18))^2$. Transformed confidence intervals can obtained in a similar fashion, by substituting the upper and lower bounds of the CI on the transformed scale for $\hat{\beta}$.

related slowdown at the second noun when attachment was non-local, but no such effect was found on the first noun. This finding is compatible with the URM and the underspecification model. Under the assumptions of the URM, the parser first attempts to attach the RC to the first noun. In the non-local attachment condition, it fails at doing so because such an attachment is unavailable. In the other two conditions, it succeeds. Subsequently, it moves on to the next word, and the relative clause is attached to the second noun in the non-local attachment condition, at some processing cost compared to the two other conditions, where RC attachment was already completed during the reading of the first noun. A necessary assumption is that the failure to attach the RC at the first noun comes at no processing cost, i.e. that attempting to attach a relative clause and succeeding requires the same amount of time as attempting to do so and failing. Although such an assumption may appear somewhat surprising and is certainly worth further investigation, the parser has good reasons to expect a potential attachment site at the next word because of the genitive maker on the first noun. Therefore, it does not need to treat the unavailability of the first noun for RC attachment as an anomaly, and may simply move on to the next word once RC attachment has failed. Under the assumptions of the underspecification parser, relative clause attachment is delayed until the second noun (*fans*) is read. At this point RC attachment is carried out in both unambiguous conditions. The underspecification parser can explain the longer reading times in the non-local condition under the assumption that (i) non-local attachment is inherently more difficult because it involves retrieval of the last noun from memory, or (ii) the parser always attempts to carry out local attachment first, and only after failing to do so, it attaches the RC non-locally.

5.2 Summary

In this chapter, an experiment was presented which tested the predictions of the strategic underspecification model and the URM in a language with pre-nominal relative clauses. The reasoning was that the strategic underspecification model must delay RC attachment until both possible attachment sites were read. At this point, the underspecification parser must decide to underspecify in ambiguous conditions, but not in unambiguous ones. A strategic underspecification parser must be capable of such a delay because according to Swets et al. (2008), it is capable of delaying RC attachment for much longer periods of time, i.e, until the question-answering phase in the Swets et al. experiment. Furthermore, it appears plausible

to assume that a strategic underspecification parser must be willing to delay attachment if such behavior has the potential to reduce the processing effort on some occasions. A URM-parser, on the other hand, has no reason to delay attachment, and should therefore attach the RC at the first opportunity. As a consequence of these different parsing strategies, the strategic underspecification model predicts an ambiguity advantage at the second possible attachment site of the sentence, while the URM predicts no such effect. An experiment designed to test this prediction was conducted, and showed no evidence of an ambiguity advantage. Moreover, it is incompatible with an ambiguity advantage of a similar magnitude as the one found by Swets and colleagues. Thus, our findings are inconsistent with the underspecification model, but compatible with the URM. In the next chapter, evidence from German will be presented, which also favors the URM over the underspecification model.

Chapter 6

Fallible Parsing and Ambiguity Resolution

In the previous chapters of the present thesis, two competing models were introduced which aim to explain the ambiguity advantage. Both accounts were somewhat modified in order to account for the results obtained by Swets et al. (2008). Chapter five presented an experiment with pre-nominal relative clauses in Turkish, which was designed to distinguish between the URM and underspecification. The results favored the URM.

In this chapter, a response-signal paradigm experiment with German relative clauses will be presented, which will provide a further test of the URM and the strategic underspecification account. In addition, the experiment will attempt to distinguish between the two present models and another possible explanation of the ambiguity advantage, which has not been discussed so far — namely that the ambiguity advantage may be caused not by faster processing, but rather by more *successful* processing of ambiguous sentences. The underlying assumption is that on some occasions, the parser simply may fails in arriving at a proper analysis for a sentence. This may happen for a variety of reasons, such as resource limitations, or failure to retrieve the preceding parts of the sentence from memory. Assuming that such failures may occur, they should arguably slow down sentence processing as the parser attempts to recover from a parsing failure. Furthermore, assuming that such failures occur, at least part of the difference between ambiguous and unambiguous sentences may be that the former are processed more accurately.

We turn to the motivation for this idea next.

6.1 Fallibility of Parsing

Like most theories of parsing, the URM and the underspecification model implicitly assume that erroneous responses to comprehension questions are either caused by inattentiveness, or by the parser's choice of the incorrect structure, such as in garden-path sentences. However, it is usually assumed that a permissible parsing operation always succeeds once it is attempted. This means, for example, that if the parser attempts to attach low in an ambiguous sentence, a low attachment structure will always be created.

However, this assumption is an oversimplification. This is because evidence suggests that specific processes involved in parsing may be affected by failure to different degrees in different conditions. For example, McElree, Foraker, and Dyer (2003) conducted a speed-accuracy tradeoff (SAT) experiment with sentences such as (22). They asked participants to determine the acceptability of such sentences at certain pre-determined lags. The sentences (22a,c,e) are acceptable, while (22b,d,f) are unacceptable, because the verb *panic* in the latter cannot be used with an object, i.e. one cannot panic someone or something. The critical experimental manipulation concerned the distance between the verb (*relished/panicked*) and its object (*scandal*). In order to decide on the acceptability of a sentence, participants needed to retrieve the noun *scandal* from memory. Then, they needed to attempt to integrate it with the verb. In the case of *relished* this integration would succeed, rendering the sentence acceptable, while in the case of *panicked* it would not. McElree et al. (2003) found that accuracy decreased with increasing distance between argument and verb. Importantly, this effect was true even when participants were given a relatively long amount of time to respond (2.5 *sec*). The authors concluded that with increasing distance between argument and verb the probability of successfully retrieving the argument decreased, and thus made successful dependency resolution less likely. McElree et al.'s (2003) further found that retrieval speed, i.e., the speed at which processing approached asymptote was equal across conditions. McElree and colleagues interpreted this finding as evidence for a content-addressable memory system underlying sentence comprehension in which items can be accessed directly via a set of *cues*. According to them, interposing more material between argument and verb causes interference and reduces the *probability* of correct retrieval, but not the retrieval *speed*. Importantly, because McElree et al. (2003) used unambiguous sentences, failure to retrieve the dependent argument cannot be explained under the assumption that the parser sometimes created an incorrect parse, because no alternative parses were available. The failure to retrieve the argument from memory

must have resulted in a complete parsing failure.

(22) It was the scandal that ...

a/b. the celebrity *relished/*panicked.*

c/d. the model believed that the the celebrity *relished/*panicked.*

e/f. the model believed that the journalist reported the celebrity *relished/*panicked.*

In another SAT experiment, Foraker and McElree (2007) demonstrated an interaction between language and memory: They found that pronouns referring to clefted NPs are more likely to be successfully resolved than those referring to unclefted NPs. Foraker and McElree explain this finding in terms of better memory encoding for clefted NPs. This means that the linguistic prominence of a phrase appears to affect the probability of its successful retrieval and thus the chances of successful dependency resolution. In consequence, Foraker and McElree's (2007) results suggest that linguistic structure building processes can be susceptible to failure, and that the degree of that fallibility may depend on the linguistic properties of the stimulus.

Importantly, Martin and McElree (2008, 2011) and Van Dyke and McElree (2011) found that higher asymptotes in SAT experiments corresponded to shorter reading times in eye tracking replications of their SAT experiments. This is consistent with the idea that processing failures result in slowed reading. Therefore, we believe that it is possible that the ambiguity advantage is at least in part due to lower failure rates in the ambiguous condition.

In the following, we will discuss the implications of parsing fallibility for the URM and the underspecification model, as well as their predictions for aspects of completion time distributions.

6.2 Fine-grained Predictions of URM and Strategic Underspecification

6.2.1 Underspecification

The underspecification account assumes that RC attachment takes place during the reading of unambiguous sentences, but not during the reading of ambiguous sentences. Therefore, the earliest point in time when all processes related to structure-building finish must occur earlier in ambiguous sentences (when the RC is not attached) than in the unambiguous conditions (when the RC is attached). The time

difference between these points should correspond to the minimum amount of time required to complete an attachment. Figure 6.1a exemplifies this prediction in terms of the probability of having successfully processed the RC at different times, as predicted by the underspecification account. The predictions in the figure are based on the assumption that the minimum amount of time required to process the RC is $400\,ms$ in the ambiguous condition. After $400\,ms$, the probability of having successfully processed the RC departs from zero in the ambiguous condition. In the unambiguous conditions, however, that probability is still zero because processing these conditions involves carrying out RC attachment *in addition* to everything that is done in ambiguous sentences. Thus, because the parser has not yet completed everything attachment-related after $400\,ms$, the probability of having processed an unambiguous sentence departs from zero later. In this example, we assumed that RC attachment requires at least $50\,ms$, and that therefore the RC cannot be processed in less than $450\,ms$ in unambiguous conditions.

6.2.2 F-Underspecification

The underspecification account does not have any clear connection to the idea of parsing failure outlined above. Therefore, it is not clear what a strategic underspecification parser would do if the parser failed to construct a parse. One possible set of assumptions one could make is this: in order to decide whether to underspecify an ambiguity, the parser needs to retrieve all potential attachment sites from memory and check whether they actually qualify for attachment.[1] Parsing failure could result from not being able to retrieve any noun phrases that pass that later step. A potential reason for failure could be that in a sentence such as (23), on some occasions, the noun *general* fails to be retrieved due to memory problems, while the noun *daughter* is retrieved, but does not qualify for attachment due to a gender mismatch with the reflexive *himself*. When this happens, the parser has effectively failed to retrieve *any* attachment site from memory, and so it fails to construct a parse. We will call this model, which assumes that parsing may fail at due to failure to find a suitable attachment site for a relative clause F-Underspecification.

(23) The daughter of the general who scratched himself felt humiliated.

F-Underspecification predicts lower failure rates in ambiguous than in unambiguous conditions. This is because in the ambiguous condition, there are two potential

[1] For example, it could check whether reflexives such as himself/herself are in agreement with an otherwise syntactically available potential attachment site.

attachment sites, while there is only one in each of the unambiguous conditions. Assuming that the probabilities of successfully retrieving the high and low attachment site (p_{high} and p_{low}, respectively) are independent, the probability of successfully retrieving either attachment site in the ambiguous condition (p_{amb}) is given by equation 6.1. Importantly, p_{high} and p_{low} also correspond to the probabilities of successful retrieval in the high and low attachment conditions. This prediction of this model with respect to the probability of successfully completing an attachment as a function of time is illustrated in figure 6.1b: F-Underspecification retains the predictions concerning differences in minimum processing times, but in addition it predicts fewer parsing failures, and therefore a higher asymptotic accuracy in the ambiguous condition.

$$p_{amb} = 1 - (1 - p_{high})(1 - p_{low}) \tag{6.1}$$

Importantly, the prediction of a lower failure rate in the ambiguous condition does not follow from the core assumptions of underspecification. If follows from an entirely separate set of assumptions which elaborates on the the parser's operational principles and links them to an underlying fallible memory system. In consequence, the strategic underspecification account is in principle compatible with other other assumptions about the link between the memory architecture and the parser's operations. These alternative assumptions may lead to predictions which differ from those of F-Underspecification. Moreover, F-Underspecification assumes two separate mechanisms which can explain the ambiguity advantage. One is the omission of the RC attachment step in ambiguous conditions, and the other is the mechanism predicting lower failure rates in ambiguous conditions, which tend to correspond to faster reading in eye tracking experiments.

6.2.3 URM

Figure 6.2a exemplifies the URM's predictions under the assumption that RC attachment does not fail. The URM predicts that ambiguous conditions should be processed faster due to *statistical facilitation*. This means that at any given time, the probability of having completed RC attachment is the complement of the probability that neither high attachment, nor low attachment, has been completed yet. Figure 6.2a shows, for example, that the probability of successful attachment is approximately 0.7 at 600 ms in the unambiguous conditions, but 0.91 in the ambiguous condition. This is because the probability that no attachment has been completed

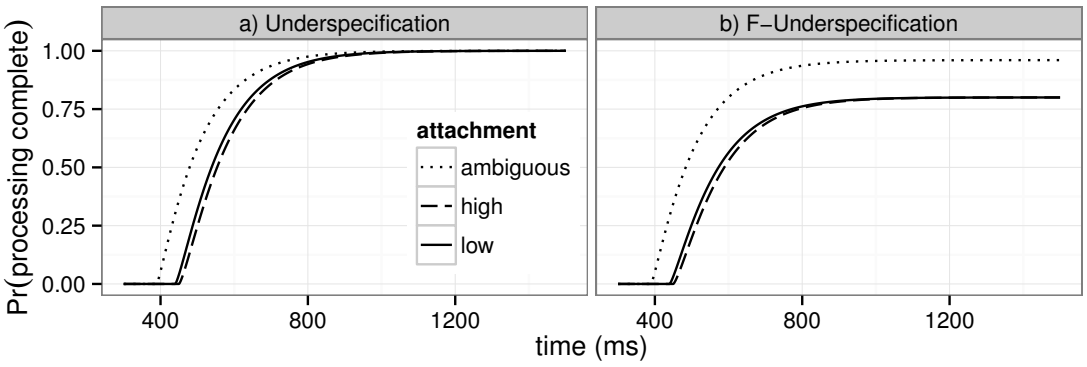


Figure 6.1: Probability of having successfully completed an attachment at a particular time as predicted by the underspecification account, F-Underspecification.

(the complementary event) is $(1 - 0.7)^2 = 0.09$. Thus, the URM predicts that the relationship between the probabilities of having successfully carried out RC attachment at any given point in time t should be as given in equation 6.2, where $p_{amb}(t)$, $p_{high}(t)$, and $p_{low}(t)$ are the probabilities of having successfully attached the RC at time t in the ambiguous, high and low attachment conditions respectively. It follows from this equation that, in contrast to both underspecification models, the URM predicts no differences in the minimum amount of time required for RC attachment. This is because the fastest attachment times in the unambiguous conditions are also the fastest completion times in the ambiguous conditions.

$$p_{amb}(t) = 1 - (1 - p_{high}(t))(1 - p_{low}(t)) \tag{6.2}$$

6.2.4 F-URM

Under the assumption that both attachment operations (i.e., high and low attachment) can fail, the URM, like F-Underspecification, predicts the failure rate in the ambiguous condition to be lower than in the unambiguous conditions. This prediction translates to a higher asymptotic success probability in ambiguous sentences. Importantly, this prediction does not follow from additional assumptions, as in the case of F-Underspecification, but from the core assumptions of the URM with no additional assumptions about the parser's operations. Figure 6.2b illustrates the predictions of the URM which directly follow from equation 6.2. F-URM, as we will call the model henceforth, does not predict a difference in the earliest completion

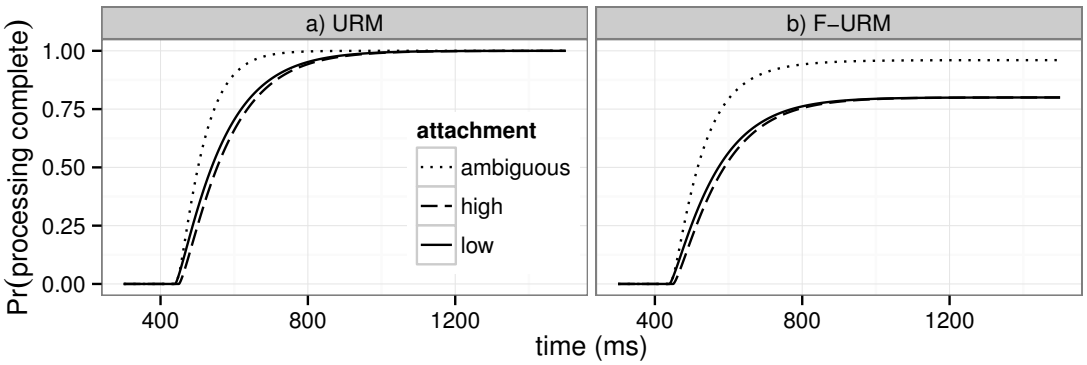


Figure 6.2: Probability of having successfully completed an attachment at a particular time as predicted by the URM, and the F-URM.

times, and it predicts that the failure rate in the ambiguous conditions should be the product of the failure rates in the unambiguous conditions.

Although a reading time or reaction time study is one way to test the predictions of the URM and underspecification concerning minimum processing time, the interpretation of such data would be complicated by the fact that some of the shortest reading times or reaction times (RTs, henceforth) may occur on trials when participants are paying little to no attention to the task. Thus, the minimum reaction time in trials on which participants did engage in processing cannot be accurately estimated. Contamination of the data by such trials may produce spurious evidence in favor of the URM or F-URM. In order to circumvent this problem, we decided to employ the so-called *response-signal paradigm* or *speed-accuracy tradeoff paradigm (SAT)* in order to test the predictions of the above models. In the following, we will present SAT, discuss the relationship between *speed-accuracy tradeoff functions (SATFs)* and completion time distributions and describe the SATF predictions for the relationship between the three attachment conditions: high, low and ambiguous.

6.3 Speed-Accuracy Tradeoff Functions and Relative Clause Attachment

It has been known since at least Pachella (1974) that accuracy can be traded off for speed in many tasks. That is, participants can choose to perform the task more accurately at the cost of lower speed, or faster at the expense of accuracy. The

function describing how accuracy depends on speed in a particular task is called a *speed-accuracy tradeoff function* (SATF). Figure 6.3 shows two typical SATFs (e.g., Wickelgren, 1977): there is an initial amount of time during which the stimulus cannot possibly have been processed yet, and so performance is at chance level. Then, as information about the stimulus starts accumulating, accuracy departs from chance. The point at which this happens corresponds to the so-called *intercept* of the SATF. The *rate* at which accuracy increases after the intercept determines the shape of the SATF. Higher rates correspond to more steeply rising SATFs. The increase of the SATF is typically negatively accelerated, which means that it rises more and more slowly as it approaches the peak performance, i.e., the *asymptote*.

A SATF provides more information about the processing in one experimental condition than mean RT and mean accuracy, such as obtained in RT tasks. This is because mean RT and mean accuracy describe one single point on a SATF, while such a point is compatible with a whole range of potential SATFs. For example, the point at the intersection in figure 6.3 is compatible with both SATFs. Which of the points on a SATF is obtained in an RT experiment depends on which speed or which accuracy the participant decides to operate at (e.g., Pachella, 1974). Importantly, an estimate of the SATF can be used to to quantify the minimum processing time (i.e., the intercept), as well as the asymptotic success probability, which is closely related to the SATF asymptote.

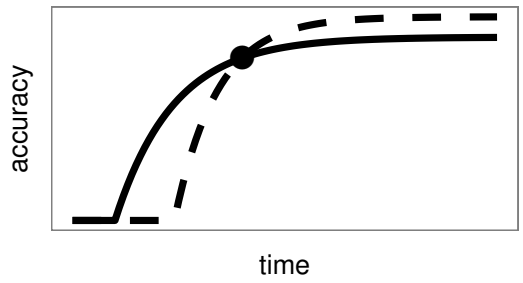


Figure 6.3: Typical speed-accuracy tradeoff functions.

In order to estimate the SATFs for the different attachment conditions in the following experiment, we used the *response-signal methodology* (Wickelgren, 1977; McElree, 1993; Liu & Smith, 2009). We presented sentences with ambiguous and unambiguous RC attachment, like those in (24), phrase by phrase, such as depicted in figure 6.4, and asked participants to categorize them as 'acceptable' or 'unacceptable'. The sentences (24a), (24b) and (24c) are 'acceptable' and correspond to the high attachment, low attachment and ambiguous conditions respectively. The

sentence in (24d) is 'unacceptable' because no attachment is possible. We enforced different attachments by manipulating the gender match between the relative pronoun and the two noun phrases in the sentence. This experimental design ensured that in order to determine whether a sentence is acceptable, participants needed to attach the relative clause. Importantly, participants were not free to choose when to respond, but were instructed to respond immediately following auditory cues. Such cues were presented at predetermined SOAs relative to the presentation of the last phrase (the RC). This procedure allowed us to estimate the accuracy of the response at different lags, and as a result, to estimate the SATF for each condition.

(24) a. HIGH

Was dachte die Managerin des Sängers, die
*What thought the manager.*FEM *of the singer.*MASC*, who.*FEM
schwieg?
was silent

b. LOW

Was dachte der Manager der Sängerin, die
*What thought the manager.*MASC *of the singer.*FEM*, who.*FEM
schwieg?
was silent

c. AMBIGUOUS

Was dachte die Managerin der Sängerin, die schwieg?
*What thought the manager.*FEM *of the singer.*FEM*, who.*FEM *was silent*

d. UNGRAMMATICAL

* Was dachte der Manager des Sängers, die
*What thought the manager.*MASC *of the singer.*MASC*, who.*FEM
schwieg?
was silent
'What did the manager of the singer who was silent think?'

Because participants' answering behavior may be biased towards 'acceptable' or 'unacceptable'-responses, we used the sensitivity measure d' (e.g., Macmillan & Creelman, 2005) as an indicator of accuracy. d' is a bias-free measure of discrimination between two types of stimuli, as is usual for experiments in this paradigm. It is computed as the difference between the z-scores of the proportion of hits and of the proportion of false alarms, where an 'acceptable' response to an acceptable sentence is considered a hit, and an 'acceptable' response to an unacceptable sentence is considered a false alarm.

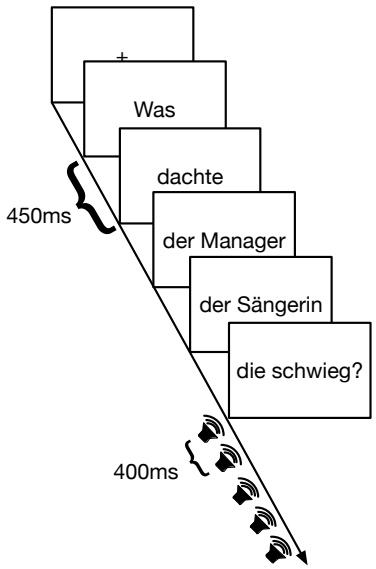


Figure 6.4: The structure of a SAT-trial.

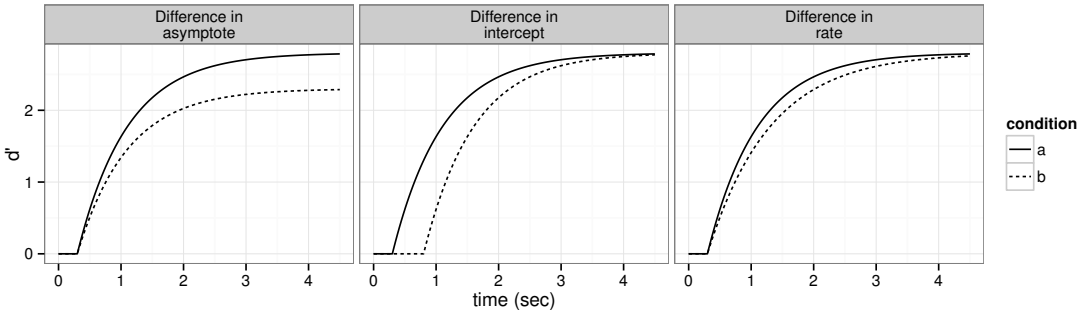


Figure 6.5: Hypothetical differences in speed-accuracy tradeoff functions.

Speed-accuracy tradeoff functions, such as those in figures 6.3 and 6.5, are well-described by a negatively accelerated shifted exponential function, such as in equation 6.3 (e.g, McElree, 1993; Liu & Smith, 2009). In this equation, λ corresponds to the asymptote in d' units, β (in ms) corresponds to the reciprocal of the rate , and specifies the amount of time required to reach approximately 63% of the asymptote. Finally, δ corresponds to the intercept (in ms).

$$d'(t) = \lambda(1 - e^{-(t-\delta)/\beta}),\ for\ t > \delta,\ \text{otherwise}\ d'(t) = 0 \tag{6.3}$$

SATFs can differ in *asymptote*, *intercept*, or *rate*, as illustrated in figure 6.5. Im-

portantly for our experiment, SATFs have a close relationship to the completion time distribution and the probability of successful processing of the process under investigation (Dosher, 1979). The intercept marks the earliest point in time at which (at least partial) information pertinent to the task becomes available. In our experiment, the intercept of the SATF corresponds to the minimum amount of time required to attach the RC. Thus, a lower intercept in one attachment condition would mean that the minimal amount of time required to process the RC in that condition is smaller. A higher asymptote in one condition corresponds to a higher probability of successfully processing the RC, i.e., building the necessary structure. Finally, the rate depends of the variability in the completion time distribution.[2]However, the rate also depends on the asymptote and so the relationship between rate and standard deviation is not easily interpretable unless asymptotes are equal. Therefore, we will focus on the predicted differences in intercepts and asymptotes.

In the following, we will describe the experiment starting with how the beforementioned predictions of the URM and the underspecification model translate into predictions about their SATFs.

6.4 Experiment 2

All four models presented above, URM, Underspecification as well their fallible counterparts predict that the asymptote in the ambiguous conditions should be either greater than or equal to the asymptotes in the unambiguous conditions. Furthermore, all models predict that the intercept in the ambiguous condition should be either smaller than or equal to the intercepts of the unambiguous conditions. The crucial parameter for distinguishing between the URM and Underspecification on the one hand, and the F-URM and F-Underspecification on the other hand is the asymptote. The F-URM and F-Underspecification predict higher asymptotes in the ambiguous conditions, while the URM and Underspecification predict all asymptotes to be equal.

[2]This is true under the assumption that RC attachment is an all-or-none process, i.e., that no partial information about RC attachment is available to the decision-making processes before attachment is completed or has failed. Such a processing mode implies that participants try to guess the right answer if RC attachment has not yet terminated, and give an informed response following completion or failure of RC attachment. Under the contrasting assumption that RC attachment is *not* all-or-none (i.e., under the assumption that participants may have *partial* information about the permissibility of a particular attachment), the rate corresponds to the average rate of information accumulation over all trials.

The crucial parameter for distinguishing between both versions of the URM on the one hand, and both versions of underspecification on the other hand is the intercept. (F-)Underspecification predicts a smaller intercept in the ambiguous condition than in any of the unambiguous conditions. (F-)URM on the other hand, predicts that the intercept of at least one unambiguous condition should be equal to the intercept of the ambiguous condition.

In sum, differences in intercept and asymptote allow us to distinguish between all four models. Importantly, the fact that (F-)URM and (F-)Underspecification make different predictions concerning the intercept follows directly from these models, and from the assumption that attachment cannot be completed in $0\,ms$ (i.e., that there is a minimum amount of time which is required to complete attachment).

6.4.1 Method

Participants

Twenty students from the University of Potsdam participated in exchange for course credit. All were native speakers of German; their age range was 18-36 years. The data of two participants was excluded from analysis because they consistently failed to identify the high attachment condition as acceptable.[3]

Procedure

To estimate the SATFs for different attachment conditions, we used the *multiple-response SAT procedure* (MR-SAT) (Foraker & McElree, 2007; Martin & McElree, 2011, 2009, 2008; McElree, 1993; Van Dyke & McElree, 2011; Wickelgren, Corbett, & Dosher, 1980). On each trial, a sentence was presented phrase by phrase in the center of the screen, as illustrated in figure 6.4. Each phrase was presented for 400 ms, with an ISI of 50 ms. 600 ms before the onset of the last phase, comprising the entire relative clause, a series of fourteen 500 Hz tones started, with an SOA of 400 ms between tones. Each tone served as a cue for the participant to classify the sentence as acceptable or unacceptable by means of pressing a button on a game-pad. Participants were told that they did not need to (but were free to) respond to the first tone, as its main purpose was to serve as a preparatory cue. The last tone sounded at 4.6 seconds after the onset of the last phrase. One half of the participants

[3]The proportion of 'acceptable' responses to high attachment sentences in both cases was less than 15%.

was requested to press a button with the index finger of their right hand to indicate the answer 'acceptable' while the other half were requested to press this button to indicate the answer 'unacceptable'.

Before taking part in the experiment, participants practiced the task. First, they were trained to indicate the orientation of an arrowhead presented on the screen by means of pressing the left or the right button on a game-pad. This procedure served to familiarize participants with the pace at which they will have to press buttons during the experiment. Next, they were trained to modulate their responses based on changes in the arrowhead's direction during the trial. This training procedure consisted of 44 trials. Lastly, they practiced the actual experimental task on 67 unrelated sentences. Participants required approximately 30 minutes for the entire training session. The actual experiment took approximately 80 minutes and consisted of 16 blocks of 33 sentences each. The first sentence of each block was a filler unrelated to the experiment. Participants were encouraged to take breaks between two blocks whenever necessary.

6.4.2 Materials

We created 32 sets of sentences like (25). Each grammatical sentence from every set was presented to each participant once. Each ungrammatical sentence was presented three times throughout the experiment in order to balance grammatical and ungrammatical sentences. We thus presented 192 grammatical experimental sentences (64 for each grammatical attachment condition) and 192 ungrammatical experimental items. The experimental sentences were intermixed with 144 additional sentences, of which one half was grammatical and the other half contained number agreement violations between subject and verb.

(25) a. HIGH, FEMININE RELATIVE PRONOUN

Was dachte die Managerin des Sängers, die
*What thought the manager.*FEM *of the singer.*MASC*, who.*FEM
schwieg?
was silent

b. HIGH, MASCULINE RELATIVE PRONOUN

Was dachte der Manager der Sängerin, der
*What thought the manager.*MASC *of the singer.*FEM*, who.*MASC
schwieg?
was silent

c. LOW, FEMININE RELATIVE PRONOUN

Was dachte der Manager der Sängerin, die
*What thought the manager.*MASC *of the singer.*FEM*, who.*FEM
schwieg?
was silent

d. LOW, MASCULINE RELATIVE PRONOUN

Was dachte die Managerin des Sängers, der
*What thought the manager.*FEM *of the singer.*MASC*, who.*MASC
schwieg?
was silent

e. AMBIGUOUS, FEMININE RELATIVE PRONOUN

Was dachte die Managerin der Sängerin, die schwieg?
*What thought the manager.*FEM *of the singer.*FEM*, who.*FEM *was silent*

f. AMBIGUOUS, MASCULINE RELATIVE PRONOUN

Was dachte der Manager des Sängers, der
*What thought the manager.*MASC *of the singer.*MASC*, who.*MASC
schwieg?
was silent

g. UNGRAMMATICAL, FEMININE RELATIVE PRONOUN

* Was dachte der Manager des Sängers, die
*What thought the manager.*MASC *of the singer.*MASC*, who.*FEM
schwieg?
was silent
'What did the manager of the singer who was silent think?'

h. UNGRAMMATICAL, MASCULINE RELATIVE PRONOUN

* Was dachte die Managerin der Sängerin, der
*What thought the manager.*FEM *of the singer.*FEM*, who.*MASC
schwieg?
was silent
'What did the manager of the singer who was silent think?'

In order to prevent participants from trying to detect patterns in the presentation sequence and thus anticipating particular kinds of stimuli, we randomized all sentences according to the following constraints: (1) We ensured that the grammaticality of a sentence could not be predicted from the grammaticality of the two preceding sentences. (2) The predictability of the experimental condition of the current sentence on the basis of the conditions of the two previous sentences was minimized. (3) The predictability of the experimental condition in which a particular item will occur next on the basis of the knowledge of its last two occurrences was

kept as low as possible. (4) The probability that one item will regularly follow or precede a particular other item was minimized as well. Given this set of constraints, we maximized the distance between repetitions of lexical material. We created one randomized list which half the participants saw in its regular order, while the other half saw it in the reverse order.

6.4.3 Data Analysis

We analyzed the data in two different ways. In one analysis, we obtained the maximum-likelihood estimates (MLEs) (Myung, 2003; Liu & Smith, 2009) of all model parameters, i.e., of the parameters λ, β and δ in equation 6.3 (repeated as 6.4) for each condition for every participant. In other words, we obtained estimates of the three parameters of equation 6.4 for every participant and every condition by finding values for them that maximized the likelihood of the data given the parameter values.

$$d'(t) = \lambda(1 - e^{-\beta(t-\delta)}),\ for\ t > \delta,\ otherwise\ d'(t) = 0 \tag{6.4}$$

In a second step, we conducted data analysis on the basis of model comparison. All candidate models were fitted to each participant's data separately in order to avoid artifacts of pooling the data of several participants (e.g., Brown & Heathcote, 2003). To this end, we determined the MLEs of the parameters λ, β and δ in equation 6.4 for each condition. However, in this analysis, we imposed various constraints on the parameter estimates in different conditions.

We compared the fit of the best-fitting models for the group and on a by-participant basis.

Candidate Models

We compared the fit of models with different constraints for the parameter values for λ, β and δ. Each model constrained the parameter values for each of the three attachment conditions in one of four ways: (1) same parameter value for all three conditions, (2) different parameter values for ambiguous and high attachment on the one hand, and low attachment on the other, (3) different parameter values for ambiguous and low attachment on the one hand, and high attachment on the other, (4) different parameter values for each attachment condition. We varied the

constraints for each parameter, resulting in a total of 64 candidate models. In the following, we will refer to each model by its numbers of asymptotes, rates and intercepts. For example, model $3\lambda-1\beta-2\delta$ refers to a model with three asymptotes, one rate and 2 intercepts. Since the contrasts 2 and 3 contrasts enforced two different parameter values for a particular parameter, we will disambiguate where necessary. Because all of our candidate models predicted the asymptote to be larger or equal in the ambiguous condition, and because they also predicted the intercept to be smaller or equal in the ambiguous condition, we constrained them to be such for all models that assume differences between conditions in these parameters.

Estimation

Search for the ML estimates of the parameter values was carried out using the *SUBPLEX* algorithm (Rowan, 1990) implemented in the *NLopt* optimization library (Johnson, 2013).[4] The log-likelihood of the data given the parameters λ', β' and δ' for each condition was computed as the sum of the log-probabilities of all responses given the SATF predicted by these parameters. The response probability for the first response on any trial was computed according to $P(R =' acceptable') = \Phi(d'(t) - c(t))$, where R is the response, and t is the time relative to the onset of the last phrase at which it was given. Φ is the cumulative distribution function of the Gaussian distribution, and c is the *criterion location* in signal detection theory (e.g., Macmillan & Creelman, 2005), which we estimated from the data.[5] All three grammatical conditions were scaled against the ungrammatical condition. For all responses following the first one, we computed conditional response probabilities (conditioned on the previous response on this trial) using Albers and Kallenberg's (1994) approximation to the probability function of the bivariate normal distribution with a correlation of ρ. The value of ρ was determined by MLE.

Model Comparison

We used the Bayesian information criterion (BIC) for model selection and the BIC approximation to the Bayes factor for inference (Wagenmakers, 2007). The BIC was computed according to equation 6.5, where $log\ L$ is the maximized log-likelihood

[4]The R-code we used for optimization is available for download at http://r-forge.r-project.org/projects/satf .

[5]We assumed that the position of the criterion location c as a function of time t is well-described by the four-parameter function $c(t) = \lambda_c(1 - e^{-\beta_c(t-\delta_c)}) + \alpha_c\ for\ t \geq \delta_c$, otherwise $c(t) = \alpha_c$, and estimated the parameters of the functions describing d' and c simultaneously.

of the data under a given model, n_{par} is the number of free parameters in the model, and n_{obs} is the number of observations. The BIC decreases with increasing log-likelihood of the data and increases with the number of free parameters. The amount of free parameter penalty depends on the amount of data (n_{obs}). The model with the smallest BIC provides the most parsimonious fit to the data, and maximizes the generalizability of a model (Pitt & Myung, 2002). Because the responses on one trial were highly correlated, we set n_{obs} to the number of trials for one participant, instead of the actual number of data points.

$$BIC = -2 \cdot log\ L + log(n_{obs}) \cdot n_{par} \tag{6.5}$$

For formal inference, we used the BIC approximation to the Bayes factor according to equation 6.6 (Wagenmakers, 2007), where BIC_1 and BIC_2 are the BIC values of the models to be compared. The Bayes factor quantifies the evidence in favor of model 1 over model 2. We combined individual participants' Bayes factors for each comparison into group Bayes factors (GBF), as suggested by Stephan and Penny (2006). The GBF for comparison of models 1 and 2 was computed according to equation 6.7 as the product of all participants' Bayes factors for that comparison (where k is an index over participants). Like a single Bayes factor, a GBF can be interpreted as the ratio of evidence in favor of model 1 and the evidence in favor of model 2. By convention (e.g., Raftery, 1995), the evidence is considered *weak* when $1 < BF < 3$, *positive* when $3 < BF < 20$, *strong* when $20 < BF < 150$, and *very strong* positive when $BF > 150$. We chose to present *log* GBFs (lGBFs) here because the value of the GBF tended to become rather large. On a log scale, values above 1 can be considered positive evidence, and a values above 3 and 5, respectively, can be considered *strong* and *very strong* evidence.

$$BF_{12} = e^{(BIC_2 - BIC_1)/2} \tag{6.6}$$

$$GBF_{12} = \prod_k BF_{12(k)} \tag{6.7}$$

We used the Bayes factor-based comparisons for models with constrained differences between asymptotes and intercepts. We supplement these comparisons with the analysis of coefficient estimates from unconstrained fully saturated models.

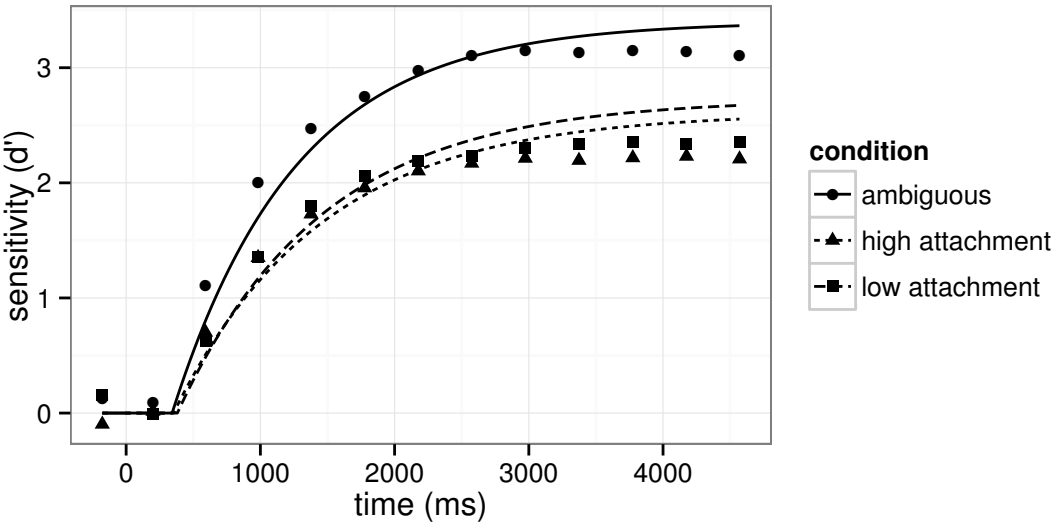


Figure 6.6: Average sensitivity (points) along with predictions of the average fully saturated model (lines).

Table 6.1: Average estimates of asymptotes, intercepts, and 1/rate for the fully saturated model ($3\lambda - 3\beta - 3\delta$). 95%-confidence intervals in brackets.

condition	asymptote	1/rate	intercept
ambiguous	3.4 [2.7; 4.1]	924 ms [595; 1253]	343 ms [233; 452]
high attachment	2.6 [2.0; 3.2]	1101 ms [541; 1661]	353 ms [214; 492]
low attachment	2.7 [2.1; 3.4]	1076 ms [628; 1523]	381 ms [265; 498]

6.4.4 Results

Parameter Estimates

Figure 6.6 shows average d' at different lags, along with the predictions of the average fully saturated ($3\lambda - 3\beta - 3\delta$) model. Table 6.1 shows the average parameter estimates for the fully saturated model, along with confidence intervals. According to the results of fully saturated model fits, the average asymptote in the ambiguous condition was higher than in the unambiguous conditions. The intercepts in the ambiguous and the high attachment conditions were approximately equal, while the intercept in the low attachment conditions was larger. The time required to reach approximately 63% of the asymptote (i.e., 1/rate) was slightly shorter in the ambiguous condition than in both unambiguous conditions.

Figure 6.7 shows the by-participant estimates of the asymptote differences between ambiguous and unambiguous conditions as well as 95%-confidence intervals (CI) for these estimates. According to the CIs, the asymptote differences between ambiguous and unambiguous conditions were significantly smaller than zero. This means that asymptotes were significantly lower in both unambiguous conditions than in the ambiguous conditions. In other words, participants were more accurate at classifying ambiguous sentences as grammatical than unambiguous sentences.

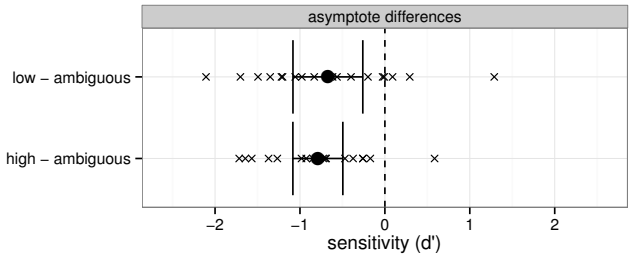


Figure 6.7: By-participant estimates of asymptote differences between ambiguous and unambiguous conditions. Crosses represent individual participants' estimates, circles represent grand means, bars show 95% confidence intervals.

Figure 6.8 shows the by-participant estimates of the intercept differences between ambiguous and unambiguous conditions as well as 95%-CIs for these estimates. Both CIs do not show a significant difference between intercepts in ambiguous and unambiguous conditions, and in fact, the CIs as well as the estimates are centered around zero. The CI for the intercept difference between ambiguous and low attachment conditions was $[-57\,ms;\ 135\,ms]$, and the CI for the difference between ambiguous and high attachment conditions was $[-64\,ms;\ 86\,ms]$.

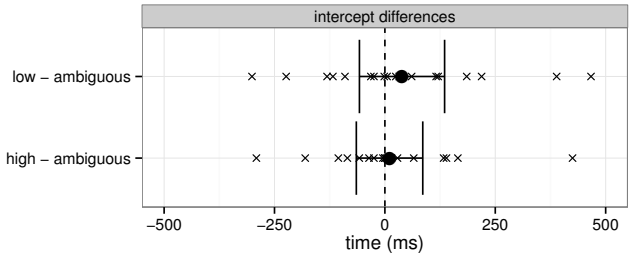


Figure 6.8: By-participant estimates of intercept differences between ambiguous and unambiguous conditions. Crosses represent individual participants' estimates, circles represent grand means, bars show 95% confidence intervals.

Figure 6.9 shows the by-participant estimates of the 1/rate differences between

ambiguous and unambiguous conditions as well as 95%-confidence intervals (CI) for these estimates. It shows that the differences in 1/rate between ambiguous and unambiguous conditions are not significant.

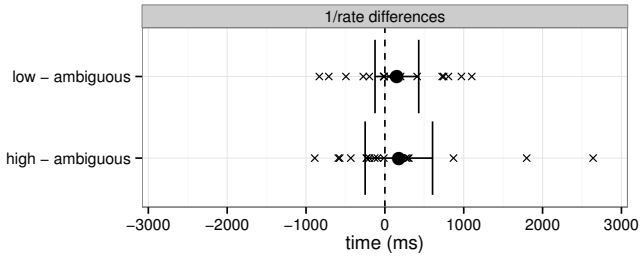


Figure 6.9: By-participant estimates of 1/rate differences between ambiguous and unambiguous conditions. Crosses represent individual participants' estimates, circles represent grand means, bars show 95% confidence intervals.

Group Model Comparison

Table 6.2 shows the results of the model comparison. The first column shows the rank of the model according to the average of its BIC across participants. The next 6 columns show the mean estimated differences between the ambiguous and each of the unambiguous conditions, where applicable. For example, the highest-ranked model (rank 1) was $3\lambda - 1\beta - 2\delta$, and therefore no estimates of 1/rate differences are presented in the first row of table 6.3. In the same table, $\Delta logL$ and $\Delta\overline{BIC}$ denote the difference in log-likelihood and average BIC between the model in that row and the highest-ranking model. The column $lGBF_1$ shows the log group Bayes factor quantifying evidence in favor of the highest-ranking model, compared to the model in the current row.

The highest-ranked model according to mean BIC was $3\lambda - 1\beta - 2\delta$, a model with three asymptotes and a higher intercept for the low attachment condition. The best model with less than three asymptotes was ranked eighth. A log group Bayes factor of 56 provided very strong evidence against the highest-ranked model with fewer than three asymptotes, and in favor of the $3\lambda - 1\beta - 2\delta$ model. This finding is in agreement with the pattern of by-participant estimates in figure 6.7.

The highest-ranked model assuming a larger intercept (in the high-attachment condition) was ranked fifth. A log group Bayes factor of 44 provided very strong evidence against this model. The evidence against intercept differences between ambiguous and high-attachment conditions agrees with the pattern of by-participant

Table 6.2: Best-ranked models according BIC based on their fits to individual participants data: ΔLL is the model's log-likelihood minus the log-likelihood of the model selected by BIC. $\Delta\overline{BIC}$ is the model's average BIC minus the minimum average BIC.

	asymptote		1/rate		intercept					
rank	high	low	high	low	high	low	$\Delta logL$	$\Delta\overline{BIC}$	$lGBF_1$	n_{par}
1.	-0.9	-0.8				77	0	0	0	11
2.	-0.9	-0.8					-77	3	23	10
3.	-0.9	-0.8	-41			77	14	4	40	12
4.	-0.9	-0.8		32		73	14	4	40	12
5.	-0.9	-0.8			71	105	9	5	44	12
6.	-0.9	-0.8	-6				-46	5	46	11
7.	-0.9	-0.7		117			-47	5	47	11
8.	-0.6			640			-109	6	56	10
9.	-0.9	-0.8			35		-72	8	72	11
10.	-0.4						-182	8	75	9

estimates in figure 6.7. However, the evidence in favor of a higher intercept in the low attachment condition, is surprising given the confidence intervals in figure 6.7.

According to mean BIC and the Bayes factor approximation, there was strong evidence against a difference in rates ($lGBF = 40$), with the highest-ranked model assuming such differences ranking third.

By-Participant Model Comparisons

In addition, in order to ensure that our results were not due to outliers, we determined the number of participants for which each contrast-parameter combination was preferred for it on the basis of the BIC.

Table 6.3 shows the results. The columns correspond to the SATF parameters: intercept, rate and asymptote. The rows correspond to different contrasts we used: for example, the first column in the first row shows for how many participants a model with equal asymptotes for all conditions had the best fit. The cell in the second row, first column shows for how many participants a model with a separate asymptote was preferred. The last row shows for how many participants a three-asymptote model was chosen. In each cell, the number in brackets indicates for how many participants this contrast was selected with a Bayes factor of 3 or more.

Table 6.3: Number of participants for which particular parameterizations were selected on the basis of the BIC. (Number of participants with a preference supported by a Bayes factor of 3 or more in brackets).

contrast	asymptote	rate	intercept
all equal	7 (4)	13 (9)	14 (13)
high different	4 (2)	1 (1)	0
low different	1 (1)	4 (1)	4 (3)
high and low different	6 (5)	0	0

The asymptote column of table 6.3 shows that not all of our participants were best fit by a three-asymptotes model – this was the case for only six out of 18 participants. However, the data of 11 participants provided sufficient evidence for lower asymptotes in at least one of the unambiguous conditions. It stands to reason that the asymptote differences for the remaining seven participants' were very small or had the opposite sign. However, for the group as a whole, the three-asymptote model provided the best fit.

The intercept column of table 6.3 shows that a larger intercept in the low attachment condition was preferred for four out of 18 participants, with a Bayes factor of more than three for three of them. This suggests that our finding of a larger intercept in the group model selection analysis is largely due to three participants providing very strong evidence for a larger intercept in the low attachment condition. Importantly, there was no indication of intercept differences between ambiguous and high attachment conditions.

Furthermore, the data of 13 out of 18 participants was best fit without any differences in rates. This finding is in close agreement with the fact that models that assume intercept differences between the ambiguous and the high attachment conditions were not preferred for any of our participants.

6.4.5 Discussion

Our two main findings in the experiment were (1) lower asymptotes in the unambiguous conditions, and (2) no intercept differences between ambiguous and high attachment conditions.

We found that, for most of our participants, asymptotes in at least one the unambiguous conditions tended to be lower than in the ambiguous condition. On the whole, our participants were best fit by a model assuming lower asymptotes in both unambiguous conditions. This finding is incompatible with the non-fallible versions

of the URM and Underspecification because they predict equal error rates in all attachment conditions. Thus, they speak in favor of F-URM and F-Underspecification. This is because, if all acceptability judgment errors were caused by factors unrelated to parsing, all experimental conditions should be affected to the same extent, and so asymptotes across experimental conditions should be equal. Thus, the finding of higher asymptotes in the ambiguous condition is compatible only with the fallible models F-URM and F-Underspecification.

But how can the fallible models explain the individual differences in asymptotes suggested by the finding that the data of five participants data were best fit by two-asymptote models (three with a Bayes factor > 3)? Meanwhile, another seven participants' data were best fit a one-asymptote model (four with a Bayes factor > 3).

F-URM and F-Underspecification can explain this behavior because they subsume their respective non-fallible counterparts. Thus, for participants best fit by two-asymptote models we have to conclude that one of the RC attachment operations is virtually infallible. For four out of five participants, that appears to be the low attachment operation. For participants best fit by a one-asymptote model we have to conclude that parsing did not fail at all in our experimental scenario, and that all errors were unrelated to parsing. While the between-participant variability in asymptote differences may deserve further attention in future research, it appears safe to conclude that the group as a whole provides evidence for fallible models of parsing.

Our second main finding was that of no intercept differences whatsoever for 14 out of 18 participants, and no differences in intercepts between the ambiguous and the high attachment condition for all our participants (16 out of 18 with a Bayes factor > 3). The confidence intervals for intercept differences were $[-57\,ms; 135\,ms]$ for the difference between ambiguous and low attachment conditions and $[-64\,ms; 86\,ms]$ for the difference between ambiguous and high attachment conditions. This finding is incompatible with (F-)Underspecification to the extent that our experimental paradigm was sensitive enough to measure it. We interpret this finding as tentative evidence against both underspecification models, because they assume that RC attachment takes place only in unambiguous sentences and thus predict an intercept difference between ambiguous and *both* unambiguous conditions, corresponding to the minimum time required to complete an RC attachment. Therefore, evidence against a difference in intercepts constitutes evidence against both underspecification models.

Because the relationship between rates on a percent-scale and on d'-scale is not straightforwardly interpretable in the presence of asymptote differences, and because the confidence intervals for $1/rate$ estimates were fairly wide, we will abstain from interpreting this portion of the results.

6.5 Models 4 and 5: Testing an Attention-based Explanation for Asymptote Differences

We found strong evidence for lower asymptotes in the unambiguous conditions. This finding does not appear compatible with the non-fallible versions of the URM and Underspecification. However, there is an alternative explanation: Participants might sometimes become distracted, and not pay attention to parts of the sentence being presented on the screen. Due to the nature of our presentation mode (RSVP), they are unable to revisit the parts of the sentence that they missed, and may have to resort to guessing whether the noun phrase they missed was a potential attachment site. Because ambiguous sentences provide two potential attachment sites, the probability of overlooking the presence of one attachment site is substantially smaller than in unambiguous sentences. Thus, the *brief distractions account* (as we will call it henceforth) can explain the higher accuracy in the ambiguous condition. Indeed, the substantial percentage of 'acceptable' responses to ungrammatical sentences shown in table 6.4 suggests that participants do get distracted regularly. This is because none of the models of ambiguity processing specify a mechanism by which ungrammatical sentences can be misperceived as grammatical.

Table 6.4: Average proportions of 'acceptable' responses at the latest lag. (SEs in brackets)

ambiguous	high	low	none
0.91 (0.07)	0.75 (0.1)	0.8 (0.09)	0.11 (0.07)

An alternative explanation for these incorrect 'acceptable' responses is that participants entirely 'zone out' on some trials, and have to resort to guessing the correct response. We will call this the *long distractions account*. The difference between the two accounts is that according to the *long distractions account*, the probability of misclassifying an acceptable sentence as 'unacceptable' should be equal in all conditions, while according to the *brief distractions account*, unambiguous sentences should be misclassified more often, since participants are more likely to not pay

attention to the only possible attachment site and then guess its gender incorrectly than to do so for *two* possible attachment sites.

In order to rule out alternative explanation of our finding regarding differences in asymptotes, we tested the quantitative predictions of both models by examining their fit to the proportions of 'acceptable'-responses at the latest lag (4.5 seconds after the onset of the last phrase). We did so under the assumption that the performance at such late lags is near-asymptotic. We did so in order to test the hypothesis that brief attentional lapses may be able to explain the obtained asymptote differences as well as the percentage of incorrect 'acceptable' responses to unacceptable sentences. If this were so, we would have to conclude that the URM, and not the F-URM, provides the most parsimonious account of the data. However, if the percentage of correct responses cannot be explained by attention alone, we would have to conclude that the difference in asymptotes between ambiguous and unambiguous conditions can only be explained by parse failures.

We will turn to the quantitative predictions of both models of attentional lapses next.

Long Distractions

Under the long distractions account, we assumed that participants paid attention to the stimulus with a probability of q_A, which was equal across conditions. On inattentive trials, which occurred with a probability of $1 - q_A$, participants were assumed to have been distracted and thus resorted to guessing by responding 'acceptable' with a probability of p_Y and 'unacceptable' with a probability of $1 - q_Y$. Consequently, the asymptotic probability of responding 'acceptable' in condition C is given by equation 6.8, where p_{C} is the probability of successfully attaching the RC in condition C, and can be computed according to equation 6.1, as predicted by F-URM and F-Underspecification. According to URM and Underspecification, $p_{\mathrm{C}} = 1$ in all conditions, whereas according to F-URM and F-Underspecification models, $p_{amb} = 1 - (1 - p_{high})(1 - p_{low})$, according to equations 6.1 and 6.2.

$$p_{\textsc{acceptable}}(\mathrm{C}) = q_A \cdot p_{\mathrm{C}} + (1 - q_A) \cdot q_Y \tag{6.8}$$

Brief Distractions

Under the brief distractions account, we assumed that participants sometimes did not pay attention to one of the phrases appearing on the screen. When this happened, they tried to guess whether that missed phrase is a potential attachment site for the relative clause. We assumed that the probability that a phrase will either be attended to, or at least guessed correctly, was q_{AG}. In consequence, the probability that participants will miss a phrase and then guess it incorrectly is $1 - q_{AG}$.

Importantly, missing a noun phrase and guessing it incorrectly amounts to mistaking one experimental condition for another condition. For example, the probability of correctly identifying an ambiguous sentence as ambiguous equals the probability of paying attention to both noun phrases or correctly guessing them $(q_{AG})^2$. In the remaining cases, the ambiguous condition is mistaken for another. For example, the probability of erroneously identifying the ungrammatical condition as ambiguous equals the probability of not paying attention to either noun phrase and then guessing both incorrectly, i.e., $(1 - q_{AG})^2$. The probabilities of perceiving one condition as another can be computed accordingly.

Under these assumptions, the probability of responding 'acceptable' in condition C is given by equation 6.9, where $r_{amb}(\text{C}), r_{high}(\text{C})$ and $r_{low}(\text{C})$ are the probabilities of classifying a sentence from condition C as a sentence from the ambiguous, high attachment or low attachment conditions respectively.

$$p_{\text{ACCEPTABLE}}(\text{C}) = r_{amb}(\text{C}) \cdot p_{amb} + r_{high}(\text{C}) \cdot p_{high} + r_{low}(\text{C}) \cdot p_{low} \tag{6.9}$$

Method

In order to decide between the two accounts of attention loss, we fit them to each participant's response accuracies at the latest response lag, under the assumption that the accuracy after 4.5 seconds is near-asymptotic. To find parameter values for the free parameters which maximized the log-likelihood, we used the SIMPLEX algorithm for (Nelder & Mead, 1965) for optimization.

We combined each of the models of attention loss with two different sets of assumptions about the underlying failure probabilities: Under the *equal asymptotes* assumption, we set all probabilities of parsing failure to zero, as predicted by the URM and Underspecification. Under the *different asymptotes* assumption, we constrained the failure probability in the ambiguous condition to be the product of the

failure probabilities in the two unambiguous conditions, as predicted by F-URM and F-Underspecification according to equations 6.1 and 6.2. If the equal asymptote models can be demonstrated to provide the most parsimonious account to the data, we would have shown that our finding of asymptote differences in the SAT-experiment can be explained by attentional lapses. If the different asymptotes model can account for the data better, we would have to conclude that our data indeed do provide evidence for fallible parsing.

Results and Discussion

Table 6.5 summarizes the results of the model fits. The summary of each model fit shows the log-likelihood differences between the models ($\Delta logL$), the average BIC ($\Delta\overline{BIC}$), the number of participants for whom this model was deemed the best on the basis of the BIC ($n_{selected}$), and the number of free parameters for every participant in this model (n_{par}). Models assuming equal asymptotes (i.e., that all differences in asymptotes are attention-related) are in left panel of table 6.5, while models that higher proportions of parsing failures in the unambiguous conditions are in the right panel.

Our results show that all models that assume that the only differences in asymptotes are attention-related (table 6.5, left column), provided a much poorer fit to the data in terms of log-likelihood and average BIC. This finding appears generalizable, as the more complex models, which assume attention-unrelated asymptote differences (table 6.5, right column), were preferred for 13 out of 18 participants, based on BIC model selection.

This finding shows that the brief distractions account assuming no parsing-related failures cannot explain the asymptote differences obtained in the experiment. This means that these differences must be due to parsing and not due to attentional drifts.

Unfortunately, table 6.5 also shows that the present data does not allow us to distinguish between different accounts of attention loss. This is because the two different attention accounts with asymptote differences were selected for almost equal numbers of participants, and because differences in BIC and log-likelihood are relatively small. However, our data is sufficient to rule out attention-based explanations of asymptote differences. While it may be possible that attention-related mechanism contribute towards the difference in asymptotes found in our experiment, it can clearly not account for it in full.

Table 6.5: Log-likelihood and BIC of models of attention loss fit to the response accuracies at the latest lag. All models follow equations 6.8 and 6.9.

	equal asymptotes				different asymptotes			
	$\Delta logL$	$\Delta\overline{BIC}$	$n_{selected}$	n_{par}	$\Delta logL$	$\Delta\overline{BIC}$	$n_{selected}$	n_{par}
long distractions	-152	14	2	2	0	0	7	4
brief distractions	-540	57	0	2	-13	1	6	4
long and brief distractions	-168	17	3	3				

6.6 General Discussion

We have discussed two existing explanations of the ambiguity advantage: the Unrestricted Race Model proposed by van Gompel et al. (2000) and the underspecification model proposed by Swets et al. (2008). We showed that while all candidate models make the same qualitative predictions concerning mean reading times, their predictions differ with respect to completion time distributions, and therefore with respect to SATFs. We then discussed the implications of a fallible memory system on the SATF predictions of both theories. With regard to parsing fallibility, we instantiated two versions each of the two models, yielding four candidate models. We then presented data from an SAT-experiment, which provided evidence in favor lower asymptotic accuracy and against lower intercepts in the unambiguous conditions. By modeling the effect of potential attention-related confounds we were able to rule out alternative explanations of our finding.

The finding regarding asymptotes suggests that parsing, just like memory retrieval is subject to failure. Because ambiguous sentences are compatible with two different structures, parsing is significantly less likely to ultimately fail in such sentences. Thus our results allowed us to confidently rule out both models that assume no attachment failure, i.e., URM and Underspecification, because they were incompatible with our finding that asymptotes in the ambiguous condition were reliably higher than in unambiguous conditions. Under the assumption that parsing is fallible, the F-URM actually predicts lower failure rates in ambiguous conditions, because this follows from its fundamental assumptions. The F-Underspecification model is also compatible with this finding, but it does not follow from its core assumptions, unlike in the F-URM.

Our finding regarding intercepts speaks against the F-Underspecification model. This is because the omission of one processing step on underspecification trials (i.e., RC attachment) predicts that the minimal amount of time required to process a relative clause should be less in ambiguous sentences than in both of their unambiguous counterparts. However, none of the 18 participants showed evidence for such a difference. Admittedly, this finding may not constitute strong evidence against underspecification, because we do not know the magnitude of the difference predicted by the underspecification models. It is possible that the minimum time to complete RC attachment is relatively short. For example, if the true intercept difference was only $20\,ms$, a three-intercept model would not necessarily provide a substantial improvement in log-likelihood over a one-intercept model. The magnitude of this difference would depend on the number of data points close to the intercept. Thus, model selection on the basis of BIC-approximation to the Bayes factor could prefer the one-intercept model on the grounds of parsimony even if there was a small intercept difference in reality. In future research, the sensitivity of our method to small intercept differences may be improved by using adaptive experimentation (e.g., Myung & Pitt, 2009) in order to increase the number of informative data points around the intercept.

Although our findings concerning intercepts alone do not constitute strong evidence against F-Underspecification, there is an additional argument against it: we found evidence for a higher asymptote in the ambiguous condition. Because higher asymptotes correspond to faster reading, the ambiguity advantage can be explained by the very same mechanism which explains the asymptote differences. The higher probability of parsing failures in unambiguous sentences, which the underspecification model needs to assume in order to account for our findings is entirely sufficient within the underspecification framework in order to account for the ambiguity advantage in reading. Thus, in the absence of mode direct evidence in favor of the parser's ability to underspecify we see no reason to assume that it may be capable of such behavior. Importantly, this argument does not apply to the F-URM, because the prediction of a race as well as the prediction of higher asymptotes in ambiguous conditions follow from the same set of assumptions according to the URM, while in the strategic underspecification model, they the two effects are predicted by entirely separate components of the sentence comprehension system.

6.7 Summary

In this chapter, an SAT experiment was presented, which showed (a) that unambiguous sentences are processed less accurately than their ambiguous counterparts, and (b) that minimal amount of time required to process an ambiguous sentence is not smaller that required to process an unambiguous sentence. The present findings concerning asymptote and intercept differences, as well as the results of the Turkish experiment taken together suggest that the URM is a more parsimonious account of the ambiguity advantage and the present data counter Swets et al.'s (2008) claim.

In the following chapter, a further modification of the URM will be presented, which is susceptible to task-demands, and its quantitative predictions will be tested experimentally.

Chapter 7

Task-Dependence of Disambiguation

In the previous two chapters, we provided evidence against the strategic underspecification account by Swets et al. (2008). We argued that, although it is in principle compatible with present results, many additional assumptions are required to accommodate the full range of findings.

Previously, it was also shown that both of Swets et al.'s core findings do not provide unequivocal evidence for task-dependence of the disambiguation strategy: the apparent absence of an ambiguity advantage in the RC question condition, as well as findings concerning differences in question-answering times, are both, in principle, compatible with the URM. Thus, SDCF's data do not provide conclusive evidence for the influence of task demands on the treatment of ambiguities because these findings are compatible with the URM.

However, there is a stronger test of the influence of task demands on parsing: when the task requires *both* readings of an ambiguous sentence to be computed, a parser which is sensitive to task demands should predict an ambiguity *disadvantage*. Although this situation was not explicitly discussed by Swets et al. (2008), we believe that it is in the spirit of Swets et al.'s proposal that the parser should compute several readings of an ambiguous sentence if the task demands require it to. Thus, ambiguous sentences should be read more slowly, because computing two readings must be more costly than computing one. Importantly, an extension of SDCF's proposal must assume that parses are computed successively. In other words: The parser first attaches high, and then low, or vice versa. It needs to assume serial computation because SDCF's explanation for the ambiguity advantage is underspecification. If it also assumed parallel processing of ambiguities, it would be equivalent to the URM. Thus, the assumption of parallel computation is incompatible with SDCF's

proposal.

Importantly though, the URM can also be extended to account for the influence of task demands. To illustrate how the consideration of task demands can be implemented in a parallel, obligatory attachment model, we will next present an extension of the URM, the behavior of which is influenced by task demands. The critical difference between SDCF's model and ours is that the computation of several parses proceeds in parallel in our model, but serially in SDCF's model.

7.1 A Multiple-Channel Model of Ambiguity Resolution

Like the URM, our proposed model, the *stochastic multiple-channel model of ambiguity resolution (SMCM)* is an obligatory attachment model, which stipulates that when the processor encounters an ambiguity, it starts building all permissible parses simultaneously as soon as possible. We assume that the earliest point at which this is possible is as soon as the constituents between which a dependency has to be established have been fully processed. In the case of RC attachment, this happens at the end of the relative clause. Each structure-building process can be considered a separate processing channel. But while the URM assumes a fixed *stopping rule* (e.g., Townsend & Colonius, 1997), according to which all structure-building terminates as soon as one permissible structure has been built, the SMCM stipulates that the stopping rule is determined by task demands: if the task does not require more than one permissible RC attachment to be constructed, the parser stops after one processing channel has terminated, i.e., after one attachment has been computed. Such a system is said to be *parallel first-terminating*, following Colonius and Vorberg's (1994) terminology. If the task requires access to all available RC attachment options, the parser may choose to wait for all permissible structures to be built. Such a system is said to be *parallel exhaustive*.[1] The SMCM stipulates that the *stopping rule* is task-dependent, and that readers, in an effort to minimize reading time, prefer a first-terminating rule, unless the task suggests that the exhaustive rule should be used. The exhaustive strategy may be preferred when the reader is aware that the sentence may have several meanings and either wants to (a) pursue both possible parses, or (b) wants to select the one with the most felicitous reading. When the first-terminating stopping rule is used, the SMCM is equivalent to the URM and predicts an ambiguity advantage (cf. fig. 7.1, upper panel), but when the exhaustive stopping rule is applied, the predictions reverse such that an ambiguity

disadvantage should be observed (cf. fig. 7.1, lower panel). The SMCM with an exhaustive stopping rule makes this prediction because the probability that one of two attachment processes will finish relatively late is bigger than the probability that one particular process will finish late. The fact that the parser waits for both attachments to be computed in the ambiguous condition, as opposed to only one particular process in unambiguous conditions, will lead to more instances of long completion times in the former case. Thus, the mean reading time is predicted to be longer in the ambiguous condition.

Importantly, the SMCM makes the same qualitative predictions as a serial model, i.e., it predicts an ambiguity disadvantage when both readings have to be computed. But while serial models are fairly unconstrained concerning the quantitative relationship between completion times in ambiguous and unambiguous conditions, the SMCM makes very precise predictions given two further assumptions:[2]

1. The completion times of the structure-building processes are *statistically independent*. This means that the speed of constructing one particular structure does not depend on the speed of any other structure-building process. In other words: the completion times of the processing channels are uncorrelated.

2. The speed of a processing channel does not depend on whether another channel is active or not. This means, for instance, that making a high attachment takes a fixed amount of time (on average), whether low attachment is permissible or not. As Townsend and Honey (2007) point out, this assumption of *context invariance* is the theoretical link that justifies the comparison of data in ambiguous conditions with data in unambiguous conditions.

Because Swets et al.'s findings are compatible with the URM, they are also compatible with the SMCM employing the first-terminating stopping rule. There is reason to believe that the design of SDCF's experiment encouraged the use of the first-terminating stopping rule in all question conditions: In the superficial and the occasional questions conditions, RC attachment was not required to answer the questions correctly, which is why the parser chose the first-terminating stopping rule, in order to minimize computational effort. In the RC question conditions on the other hand, the phrasing of the questions may have motivated participants to

[1] Clearly, other stopping rules are theoretically possible. For example, the parser may wait for a fixed number of processes to finish. However, for our present purposes we distinguish only between the first-terminating and the exhaustive stopping rules.

[2] These assumptions also render it a *parallel model with independent processing channels* (Townsend & Ashby, 1983).

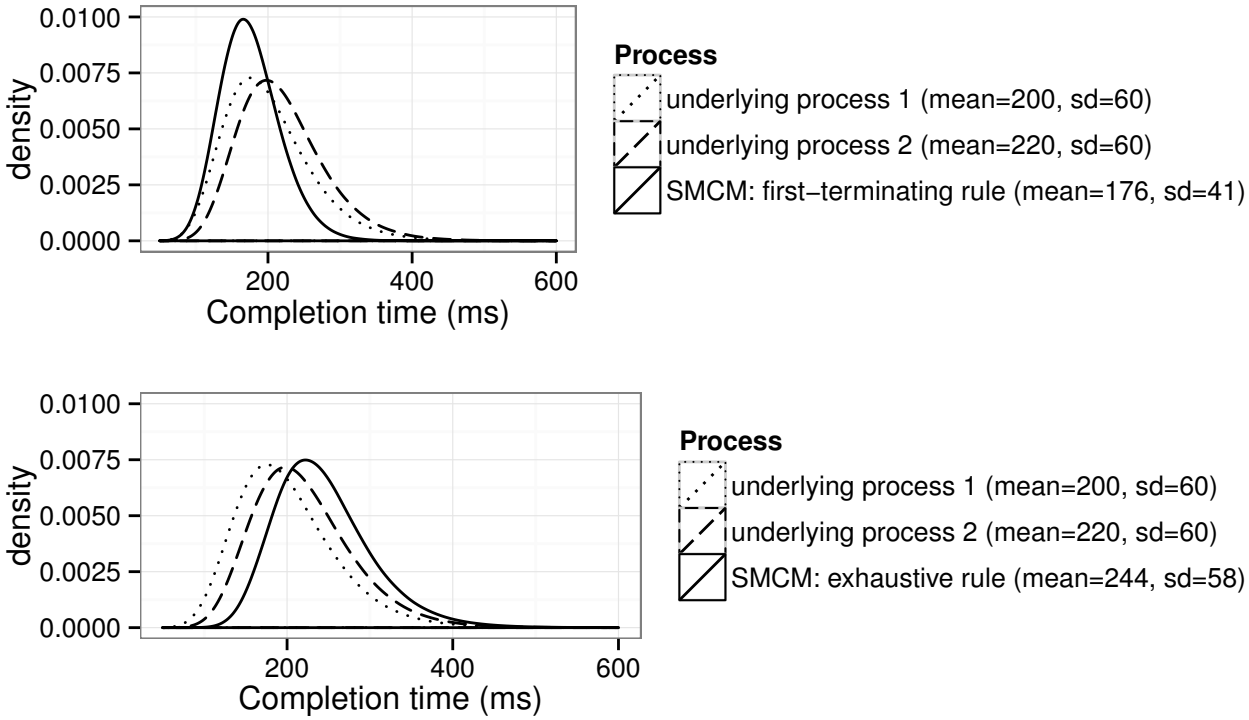


Figure 7.1: Simulated completion time distributions predicted by SMCM and the underlying attachment processes. The RTs of both racing processes are assumed to be log-normally distributed, with means 200 ms and 180 ms and a standard deviation of 60 ms.
Upper panel: SMCM with a first-terminating stopping rule. The SMCM predicts shorter mean completion times than those for any of the underlying attachment processes.
Lower panel: SMCM with an exhaustive stopping rule. The SMCM predicts longer mean completion times than those for any of the underlying attachment processes.

use the first-terminating rule: Recall that participants were asked questions such as *'Did the maid scratch in public?'* after reading sentences like (26). The readers had to respond with either 'yes' or 'no', while the option *'I don't know.'* was not available. Asking such 'yes'/'no'-questions amounts to presupposing that the reader must know the correct answer. This, in turn, is only possible if the sentence is unambiguous, because given an ambiguous sentence, a reader cannot possibly know the correct answer, since it depends on the intended RC attachment. If the RC in (26) attached high, the correct answer to the above question would be 'yes', but if it attached low, the answer would be 'no'. Therefore, the questions may have led the participants to treat the sentences as if they were unambiguous, thus leading them to pursue a first-terminating strategy since waiting for a second parse only makes sense if sentences are ambiguous.

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

In the following, we will first test the idea of task-dependence of ambiguity resolution experimentally, and then test the qualitative predictions of SMCM.

7.2 Experiment 3

In our experiment, we asked participants to read German sentences of the form of (27), in which the RC attachment could be high, low or ambiguous. All experimental sentences contained an NP-of-NP complex noun phrase, followed by a relative clause which either attached unambiguously to the first noun phrase, as in (27), unambiguously attached to the second noun, as in (28), or was globally ambiguous, as in (29). Disambiguating was effected by using the fact that in German, the relative pronoun indicates coindexation with a masculine or feminine noun through gender marking. A relative pronoun in a subject relative coindexed with a masculine noun is written *der*, whereas a relative pronoun in a subject relative coindexed with a feminine noun is written *die*. Thus, global ambiguity can be induced by making both nouns masculine or both nouns feminine, and unambiguous high and low attachment can be induced by making the first (respectively, second) noun match in gender with the relative pronoun. In order to counterbalance gender across conditions, there were two versions of each condition: one with a masculine, and one with a feminine relative pronoun. Recall that in English disambiguation was manipulated by introducing a reflexive inside the relative clause, after the relative clause

verb was read; by contrast, in our experiment, attachment was disambiguated at the earliest possible point, at the relative pronoun.

(27) HIGH ATTACHMENT

a. Die Buchhalterin des Unternehmers, die ...
*The accountant.*FEM *the.*POSS *entrepreneur.*MASC *who.*FEM

b. Der Buchhalter der Unternehmerin, der ...
*The accountant.*MASC *the.*POSS *entrepreneur.*FEM *who.*MASC

... viel goldenen Schmuck hat, hat momentan Urlaub.
lots golden jewelry has, has currently holiday.

'The accountant of the entrepreneur, who has a lot of golden jewelry, is on holiday at the moment.'

(28) LOW ATTACHMENT

a. Der Buchhalter der Unternehmerin, die ...
*The accountant.*MASC *the.*POSS *entrepreneur.*FEM *who.*FEM

b. Die Buchhalterin des Unternehmers, der ...
*The accountant.*FEM *the.*POSS *entrepreneur.*MASC *who.*MASC

... viel goldenen Schmuck hat, hat momentan Urlaub.
lots golden jewelry has, has currently holiday.

'The accountant of the entrepreneur, who has a lot of golden jewelry, is on holiday at the moment.'

(29) AMBIGUOUS ATTACHMENT

a. Die Buchhalterin der Unternehmerin, die ...
*The accountant.*FEM *the.*POSS *entrepreneur.*FEM *who.*FEM

b. Der Buchhalter des Unternehmers, der ...
*The accountant.*MASC *the.*POSS *entrepreneur.*MASC *who.*MASC

... viel goldenen Schmuck hat, hat momentan Urlaub.
lots golden jewelry has, has currently holiday.

'The accountant of the entrepreneur, who has a lot of golden jewelry, is on holiday at the moment.'

After reading a sentence, participants were asked questions such as (30), in which we wanted to know if the sentence they just read *'stated or possibly implied'* a particular attachment. With questions phrased that way, the correct answer when sentences are ambiguous is always 'yes', because both attachments (high and low)

are among the possible meanings of an ambiguous sentence. Thus, in order to answer the question correctly, the parser needs to construct both readings. Therefore, we would expect readers to construct both possible parses in the ambiguous condition if disambiguation is subject to task demands. It follows that SMCM as well as SDCF's account predict an *ambiguity disadvantage* in the present experiment. Meanwhile, the URM predicts an *ambiguity advantage* or no (detectable) effect.

(30) Wurde gerade gesagt, oder möglicherweise gemeint, dass der Buchhalter
Was just now said, or possibly meant, that the accountant
viel goldenen Schmuck hat?
lots golden jewelry has.
'Did the sentence state or possibly imply that the accountant has a lot of golden jewelry?'

7.2.1 Method

Participants

Thirty-six undergraduate students from University of Potsdam, Germany participated in exchange for course credit or 7 Euros.

Procedure

The task was self-paced non-cumulative word-by-word reading. Presentation and recording was done with the Linger software package, version 2.94 by Doug Rohde. At the beginning of a trial the whole sentence appeared, masked by underscores. Participants pressed the space bar to reveal the next word. As the next word appeared, the current one was masked by underscores again. The time between keypresses was recorded as the reading time for the word. Each sentence was followed by a question, which participants had to answer with 'yes' or 'no' by pressing the corresponding button on the keyboard.

Materials

Seventy fillers were mixed with thirty-six experimental items, each implementing the six sentences in (27), (28) and (29). Every experimental sentence began with a complex noun phrase involving two human nouns and was followed by a relative clause, which always consisted of five words. The relative clause mostly denoted a

Table 7.1: The four different kinds of questions asked in the experiment.

Did the sentence state or possibly imply that, ...	
...the accountant has a lot of golden jewelry?	*(RC/NP1)*
...the entrepreneur has a lot of golden jewelry?	*(RC/NP2)*
...the accountant is on holiday leave?	*(MC/NP1)*
...the entrepreneur is on holiday leave?	*(MC/NP2)*

possessive relationship and the verb was always a form of *"to have"* (*e.g., had long hair, had a good sense of humor, had a cold*). The relative clause attached to either the first noun phrase (NP1), the second noun phrase (NP2), or either. Relative clause attachment was disambiguated by gender agreement between the relative pronoun and the antecedent. In order to counterbalance gender across different types of attachment (*low*, *high*, and *ambiguous*), each attachment was implemented in two sentences for every item, one with a masculine relative pronoun, and one with a feminine relative pronoun. All questions combined the proposition of either the relative clause (RC) or the main clause (MC) with either NP1 or NP2. This resulted in four types of questions (RC/NP1, RC/NP2, MC/NP1, MC/NP2). Examples are provided in table 7.1. The correct response was always 'yes' for MC/NP1 questions, and always 'no' for MC/NP2. The correct response to RC/NP1 questions was 'yes' in the high-attachment condition, and 'no' in the low-attachment condition. Correct responses were reversed for RC/NP2 questions. In ambiguous conditions, the correct response was always 'yes' because all questions were embedded in the sentence frame "Did the sentence state, or possibly imply that ____?" ("Wurde gerade gesagt, oder möglicherweise gemeint, dass ____?"). Two thirds of the questions were about the main clause, while one third was about the relative clause. This was done in order to avoid focusing participants' attention on the relative clause. To make each item appear in every condition across participants, six lists were initially created. Then, in order to present every sentence with both, RC and MC questions, 12 lists were created by pairing a different subset of 12 items with questions about the relative clause.

7.2.2 Question Norming Study

We conducted a questionnaire study in order to ensure that our questions had the desired effect, i.e., that they encouraged participants to construct both readings.

Twenty-four undergraduate students from the University of Potsdam participated in exchange for course credit. We mixed our thirty-six experimental items with thirty-six filler sentences, and asked participants to answer questions about them. All questions concerning experimental sentences in the questionnaire concerned relative clause attachment. Table 7.1 shows the proportions of 'yes'-responses by condition. We excluded the data of three participants, because they responded incorrectly to questions about one of the unambiguous conditions in more than 80% of the cases.

Table 7.2: Question Norming Study: Proportion of 'yes'-responses by attachment condition and question type. Standard errors in brackets.

	high attachment	low attachment	ambiguous
RC/NP1	0.91 (0.03)	0.10 (0.03)	0.80 (0.04)
RC/NP2	0.04 (0.02)	0.82 (0.04)	0.66 (0.04)

If readers were to always adopt exactly one reading, the probabilities of adopting a high attachment or a low attachment reading respectively should sum to one. Thus, the probabilities of replying 'yes' to either question type (RC/NP1 and RC/NP2) should sum to 1 if readers adopt exactly one reading on any given trial. Clearly, when no reading is adopted (e.g., due to attention loss), the probabilities will sum to less than one.

This appears to be so in the high and low attachment conditions, where the sums of percentages of 'yes'-responses are 0.95, and 0.92, respectively. This is because participants constructed the only permissible parse in these conditions. In the ambiguous condition, however, the proportion of 'yes' responses to questions about NP1 is 0.80. If participants were building only one structure, we would expect the proportion of 'yes' responses to questions about NP2 to be approximately 0.20. However, this proportion is 0.66, with a standard error 0.04. Therefore, participants cannot be building only one structure.

7.2.3 Results

Reading Times

Table 7.3 provides an overview of the mean reading times for each word position in the relative clause, and table 7.5 shows the reading times for the spill-over regions. We excluded the data of one participant who answered questions concerning filler

sentences with a near-chance accuracy (54%), while the remaining participants performed at an accuracy above 78%. We used the *lme4* package (Bates et al., 2014) in R (R Core Team, 2013) to fit linear mixed-effects models (Pinheiro & Bates, 2000; Baayen, 2008; Gelman & Hill, 2007) to the reading time data. To determine the appropriate transformation for the dependent variable, we used the Box-Cox method (Box & Cox, 1964; Venables & Ripley, 2002). The reciprocal transformation ($1/RT$) was suggested as the most appropriate transformation for all regions we conducted analyses on. Therefore, all analyses presented here are based on reciprocally transformed reading times ($-10^5/RT$). All models included fixed effects of attachment using treatment contrasts with the ambiguous condition as a baseline. We also included random intercepts for participants and items, as well as by-participant and by-item random slopes. We did not include correlations between random intercepts and random slopes, because some models produced pathological estimates of such correlations (i.e., 1 or -1) and because according to Barr et al. (2013) models without random correlations do not significantly differ from maximal models in terms of controlling the Type-I and Type-II error rates. Outlier removal was performed using a variant of the technique recommended by Baayen and Milin (2010): we iteratively removed the data point corresponding to the largest outlying residual until the model's residuals appeared approximately normal. The rationale behind our exclusion criterion was to fulfill the assumption of normality of residuals while making maximum use of the information in the data by excluding as few data points as possible and while avoiding a *one-size-fits-all* exclusion criterion, which will lead to more missing data for slower participants. For each of the reported models we had to exclude ten data points or less. For the analysis at the RC verb (the critical region), we excluded 2 values, both smaller than 182 ms. For all six models taken together, we excluded 43 values ranging from 127 ms to 2179 ms, with a median of 188 ms. In all models presented, $|t| > 2$ and $|z| > 2$ correspond to a significant effect at a significance level of .05.

Table 7.3: Mean reading times in ms for all word positions in the relative clause. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

	relative pronoun	adverb	adjective	noun	RC verb
high attachment	543 (17)	481 (12)	529 (15)	528 (13)	622 (24)
low attachment	541 (15)	497 (14)	516 (23)	548 (17)	637 (30)
ambiguous	562 (18)	486 (14)	512 (15)	544 (15)	765 (41)

Tables 7.4 and 7.6 provide the details of the analysis. We found no significant effects of attachment in reading time at the relative pronoun ($|t|$s ≤ 1.01, or the

Table 7.4: Linear mixed-effects models coefficients, their SEs, and corresponding t-values, for the analyses of reading times at the regions relative pronoun, noun phrase (adverb + adjective + noun), and RC verb.

	relative pronoun		noun phrase		verb	
	Est. (SE)	t	Est. (SE)	t	Est. (SE)	t
High-Ambiguous	-4.25 (4.21)	-1.01	0 (1.06)	0	-13.96 (4.47)	-3.12
Low-Ambiguous	-0.91 (4.32)	-0.21	0.07 (1.21)	0.06	-12.44 (5.53)	-2.25

Table 7.5: Mean reading times in ms for the spill-over regions. Within-subject standard errors in brackets (Cousineau, 2005; Morey, 2008).

	RC verb +1	RC verb +2	RC verb +3
high attachment	492 (14)	459 (11)	693 (31)
low attachment	513 (17)	516 (25)	740 (36)
ambiguous	524 (17)	503 (17)	773 (44)

Table 7.6: Linear mixed-effects models coefficients, their SEs, and corresponding t-values, for the analyses of reading times at the three spill-over regions after the RC verb.

	verb + 1		verb + 2		verb + 3	
	Est. (SE)	t	Est. (SE)	t	Est. (SE)	t
High-Ambiguous	-5.62 (3.9)	-1.44	-12.37 (3.88)	-3.19	-9.74 (4.63)	-2.1
Low-Ambiguous	0.03 (4.01)	0.01	-3.48 (4.83)	-0.72	-6.6 (5.48)	-1.2

noun phrase consisting of the three following words ($|t|$s ≤ 0.06). There were no significant differences in reading times at any of the three words making up the noun phrase either (all $|t|$s < 1.5). However, we found a significant difference at the verb: it was read more slowly in ambiguous sentences compared to low-attachment sentences ($\hat{\beta}$=-12.44, SE=5.53, t=-2.25) and high-attachment sentences ($\hat{\beta}$=-13.96, SE=4.47, t=-3.12). Furthermore, the ambiguous condition was read more slowly than the high-attachment condition at the second word after the verb ($\hat{\beta}$=-12.37, SE=3.88, t=-3.19) and the word after that ($\hat{\beta}$=-9.74, SE=4.63, t=-2.1).[3]

Accuracy

Table 7.7: Accuracy by attachment and type of question. Within-subject standard errors for proportions in brackets (Cousineau, 2005; Morey, 2008). For RC questions in the ambiguous conditions, only 'yes'-responses were considered correct.

	MC questions accuracy	RC questions accuracy
high attachment	0.96 (0.01, N=280)	0.87 (0.03, N=140)
low attachment	0.85 (0.03, N=280)	0.79 (0.04, N=140)
ambiguous	0.86 (0.03, N=280)	0.42 (0.05, N=140)

Table 7.8: Generalized linear mixed-effects models coefficients, their SEs, and corresponding z-values for the analysis of the percentage of correct answers to MC questions.

	Est. (SE)	z value
High-Ambiguous	2.18 (0.5)	4.37
Low-Ambiguous	-0.05 (0.25)	-0.19

Table 7.9: Generalized linear mixed-effects models coefficients, their SEs, and corresponding z-values for the analysis of the percentage of correct answers to RC questions.

	Est. (SE)	z value
Low-High	1.76 (0.2)	8.75

Table 7.7 provides an overview of the mean percentages of correct question responses. Here too, we used the *lme4* package (Bates et al., 2014) in R (R Core Team, 2013)

[3]The pattern of results remained when all RTs above $3000ms$ were excluded: ambiguous sentences were read more slowly than high-attachment sentences (t=-2.7) at the verb, as well on the second and third word following it (t=-3.19, and t=-2.1). The slowdown for ambiguous sentences compared to low-attachment sentences at the verb was marginally significant (t=-1.92).

to fit generalized linear mixed-effects models assuming a logit link function to the response data. Correct responses were coded as 1, incorrect responses as 0. All models included random intercepts for participants and items, as well as by-item and by-participant random slopes for all fixed effects, but no correlations between random intercepts and slopes. We analyzed the response accuracy for MC questions in all conditions and for RC questions in the unambiguous conditions; this was because only questions concerning unambiguous sentences had unequivocally correct answers, whereas responses to questions concerning sentences with an ambiguous attachment reflected a preference.

A multilevel model fit to MC questions accuracy data (cf. Table 7.8) revealed a significant effect of attachment type indicating higher accuracy in high attachment conditions than in ambiguous conditions ($z = 4.37$, $p < 0.001$). We found no significant difference between low attachment and ambiguous conditions. Another model, which was fit to RC questions accuracy data (cf. Table 7.9) revealed a significant effect of attachment type ($z = 8.75$, $p < 0.001$) indicating higher accuracy for NP1 attachment conditions. The low proportion of 'yes' responses to RC questions in the ambiguous conditions was due to a low proportion of 'yes'-responses to *both* types of questions: For questions about attachment to NP1, the proportion was 0.49 (SE=0.07, N=70), and for questions about attachment to NP2 it was 0.36 (SE=0.07, N=70).

7.2.4 Discussion

To summarize our main findings, relative clauses with ambiguous attachment, such as (29), were read more slowly than relative clauses that unambiguously attached either high or low, such as (27) or (28). This effect occurred at the verb, which was also the last word of the relative clause. This slowdown suggests that both structures, high- and low-attachment, were computed in ambiguous conditions. Most importantly, the above-mentioned slowdown at the RC verb in the ambiguous condition is unexpected under the URM account of ambiguity resolution. Its presence suggests that, at least under some task demands, there can be an ambiguity *disadvantage*. Interestingly, this effect did not occur until the last word of the relative clause, although disambiguation happened at the first word. The good-enough parsing account proposed by Swets et al. (2008) as well as the SMCM agree very well with our results, because they assume that the parsing strategy is subject to task demands. Under these accounts, the parser computes both meanings if the questions suggest that both meanings of an ambiguity should be computed. Because computing both

meanings requires more time than computing just one, reading slows down at the verb. Furthermore, the location of the slowdown suggests that the relative clause attachment operation took place after the entire relative clause was read, which is compatible with the SMCM as well as with the good-enough account. The URM, however, cannot explain this effect without additional assumptions.

In addition to the slowdown at the verb, we found a speedup in high attachment (relative to ambiguous sentences) one word after the verb (on positions verb+2 and verb+3), but no significant difference between high attachment and ambiguous sentences. There are two possible explanations for this speedup. It is possible that faster reading in the high attachment condition reflects a delayed effect of a high-attachment preference in German (Hemforth, Konieczny, & Scheepers, 2000). However, the fact that there is no such difference at the word following the relative clause renders this explanation unlikely. Thus, we believe that this effect is more likely to reflect the retrieval of the sentence subject (Lewis & Vasishth, 2005). We will assume that attaching a relative clause to a noun phrase requires its retrieval from memory, i.e., reactivation of its memory trace. According to Vasishth and Lewis (2006), retrieving a constituent increases the strength of a memory trace and thus facilitates later retrievals. Thus, reading was facilitated in the high attachment condition because only the subject noun phrase had been retrieved earlier, during the process of RC attachment, rendering it the most active noun phrase in memory and thus facilitating its later retrieval. In the ambiguous condition, however, both noun phrases are retrieved for the purposes of RC attachment, leading to similar activation levels for both. In the low attachment condition, only the embedded noun phrase is activated, and retrieval of the subject noun phrase is not facilitated.

In addition to the findings in reading times, we found that participants were significantly better at answering questions about the main clause in the high-attachment condition than in the two other conditions. A similar pattern was found for relative clause questions. These findings suggest that it may be harder to maintain a memory representation of two clauses with different subjects than to represent two clauses sharing one subject. We also found that the proportion of 'yes' responses to RC attachment questions in ambiguous sentences (i.e., 42%) was unexpectedly low given that the slowed reading of the ambiguous relative clause suggests that both RC attachments are computed. If both parses (high and low attachment) had been maintained until the question-answering phase of the trial, the correct answer should have been 'yes' in all attachment conditions. One possible explanation for this surprising finding is that participants did compute both attachments of the rel-

atives clause in accordance with the hypothesized task requirements, but retained only one of the structures. A possible reason to do so is that the parser may not be able to maintain more than one reading simultaneously, or, in other words, that parsing is serial (e.g., Frazier, 1987; Lewis, 2000).

In sum, our results suggest that readers can build two attachments (hence the slowdown in the ambiguous conditions), but that they retain only one of them. When taken together with previous findings, they show that parsing is susceptible to task demands. According to the SMCM, when question difficulty is low, processing is relatively shallow, and so structure building terminates as soon as one parse is built—for shallow processing any one attachment will suffice. When question difficulty is high, the parser waits for both structures to be built, and then possibly selects the more plausible one. Such a mechanism predicts an ambiguity advantage in the first case, and an ambiguity disadvantage in the second.

SDCF's theory can predict the same pattern, but for somewhat different reasons. While the ambiguity advantage is explained by strategic underspecification, an ambiguity disadvantage would need to be explained by the assumption that the parser builds two structures sequentially, when questions are difficult. SDCF would need to assume sequential structure-building because by rejecting the race account of the ambiguity advantage, they also reject the mechanism of simultaneous structure-building.

Hence, the crucial differences between the SMCM and SDCF's theory lie in (i) the explanation for the ambiguity advantage (underspecification vs. race) and (ii) the explanation for the ambiguity disadvantage (two successive attachment operations vs. waiting for the second of two concurrent attachment operations to finish). In spite of different mechanisms, both theories make the same qualitative predictions. However, the SMCM also makes quantitative predictions concerning the reading time in the ambiguous conditions on the basis of the reading times in the unambiguous conditions. Unfortunately, the quantitative predictions of the good-enough account are much less clear. This is so because they depend on the duration of an attachment operation, which is unknown.[4] In the following section, we will test the quantitative predictions of the SMCM.

[4]All we can assert about an attachment operation in our experiment is that its duration is estimated to be between 0 ms and 622 ms (the reading time in the faster one of the unambiguous conditions). Although a part of these 622 ms must be due to processes such as word recognition, and pressing a button, and RC attachment is therefore likely to require significantly less time than 622 ms, we have no way of estimating its duration. Therefore, the good-enough model is compatible with any ambiguity disadvantage of less than 622 ms. The model's quantitative predictions are therefore fairly unconstrained and no meaningful quantitative predictions can be derived.

7.3 Model 6: Testing the Predictions of the SMCM

The predictions of the SMCM for our experiment are more straightforward to derive than the predictions of a serial model because we do not need an estimate of the time required by an additional attachment operation. It follows from SMCM's context invariance assumption that the variability of attachment time for high and low attachment processes is equal in ambiguous and unambiguous attachment conditions. Although factors such as lexical processing must undoubtedly contribute some additional variability in RT across condition, we expect such influences to be minor, because the verb in all items was short high-frequency word ("hat" or "hatte"). Thus, we assume that the amount of attachment-unrelated variability in RT is negligible, and therefore make the simplifying assumption that for any given participant the amount of processing time contributed by attachment-unrelated factors is constant. We therefore used the RT variability in the low and high attachment conditions as estimates of the attachment completion time variability of the low and high attachment processes, respectively. Under this simplifying assumption, we can predict the reading time in the ambiguous conditions from the reading times in the unambiguous conditions in order to examine how well the SMCM performs in the light of the evidence.

7.3.1 Method

In order to simulate the predictions of the SMCM with an exhaustive stopping-rule we used the data in the unambiguous conditions to estimate the completion time distributions for the high and low attachment processes for each participant. Based on these estimates, we repeatedly generated samples of RTs predicted by SMCM for each participant. This procedure allowed us to generate a prediction for the mean RT in the ambiguous condition, as well as to quantify our uncertainty about the prediction. Quantifying this uncertainty is important because our predictions were generated on the basis of the reading times in unambiguous conditions, which are subject to sampling error. On the basis of the obtained variability of our predicted mean RT, we were able to determine a 95% confidence interval for our prediction.

In order to obtain a prediction, we proceeded in the following steps: for each participant and every unambiguous condition, we used the reading times at verb to estimate the parameters of a log-normal distribution representing that participant's attachment completion times in that condition. We then repeated the following steps 100,000 times:

1. For each participant, we randomly drew 12 pairs of reading times from the above-mentioned log-normal distributions. In each pair, one value was from the high-attachment distribution, and the other from the low-attachment distribution.

2. We simulated the predictions of the SMCM with an *exhaustive* stopping rule by averaging the *maxima* of the 12 pairs to obtain a bootstrap sample for one participant.

3. We averaged the bootstrap samples for all participants to obtain one bootstrap sample for the entire group.

We repeated the above steps in order to estimate the mean reading time predicted for the ambiguous condition and a 95% confidence interval based on variability in the predictions (due to the variability in the data in the unambiguous conditions). This technique (repeatedly sampling from a distribution in order to estimate the variability of a statistic of interest) is known as *stratified* or *blocked bootstrap resampling* (Hesterberg, Moore, Clipson, & Epstein, 2005; Wehrens, Putter, & Buydens, 2000). We also simulated the predictions of the SMCM with a first-terminating stopping rule by averaging over the minima instead of the maxima of the 12 pairs of reading times. The results of this simulation are shown in figure 7.2.

Moreover, we carried out the procedure under two other sets of assumptions, to verify that our results do not depend on the assumption that RTs are log-normally distributed: (i) non-parametric bootstrapping, and (ii) parametric bootstrapping an ex-Gaussian distribution.[5] For non-parametric bootstrapping, we sampled (with replacement) from the reading times in the unambiguous conditions, and for parametric bootstrapping under the assumption that RTs in the unambiguous conditions are distributed according to an ex-Gaussian distribution. We used an iterative algorithm (Nelder & Mead, 1965) in R (R Core Team, 2013) in order to find distribution parameters maximizing the likelihood of the data in each condition (Lacouture & Cousineau, 2008; Myung, 2003).

7.3.2 Results

Figure 7.2 shows the mean predicted reading times obtained via bootstrap resampling, along with 95% confidence intervals and the mean reading time for ambiguous

[5] Both distributions, the log-normal and the ex-Gaussian, have been used as descriptive models of reaction time (e.g., Ulrich & Miller, 1993; Van Zandt, 2002).

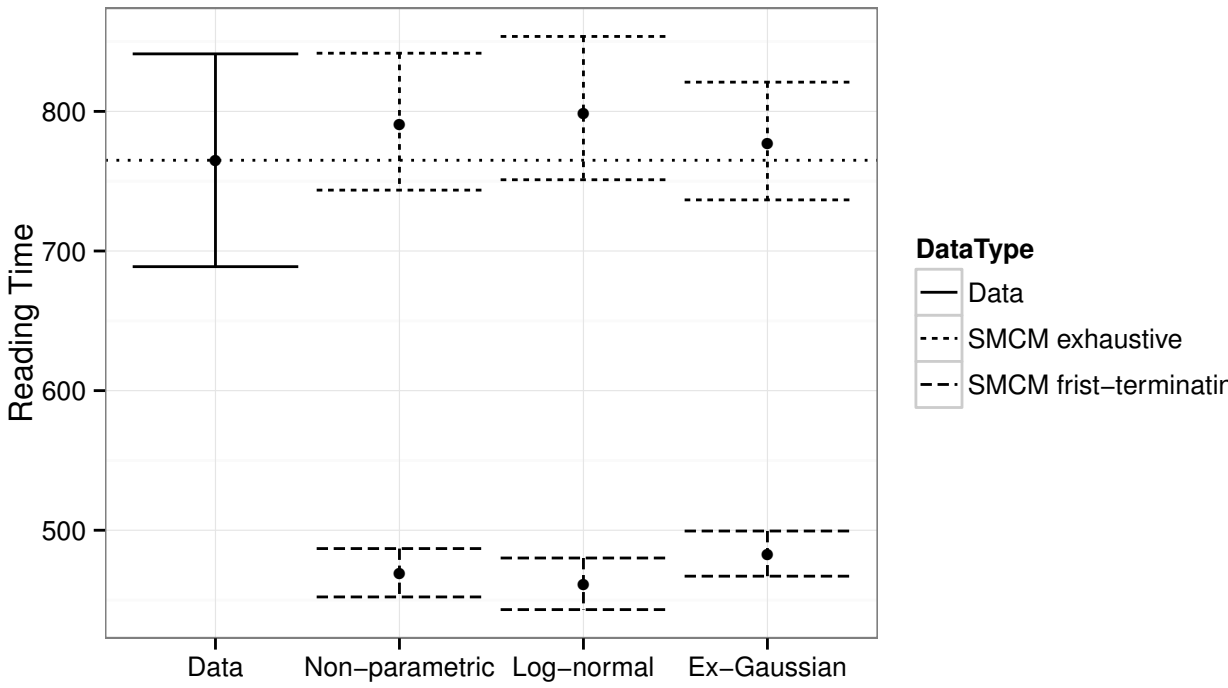


Figure 7.2: Mean reading times in the ambiguous condition and mean reading times predicted by SMCM and the associated 95% confidence intervals.
On the left, the empirical mean reading time at the verb in the ambiguous condition and its associated 95% confidence interval. On the right, 'Non-parametric', 'Log-normal' and 'Ex-Gaussian' show mean reading times predicted by the SMCM (bold points) and the associated 95% confidence intervals obtained by bootstrap resampling based on the respective distributional assumptions for reading times in the unambiguous conditions.

sentences obtained in the experiment. According to our simulations, the SMCM with an exhaustive stopping-rule predicted the reading time for ambiguous conditions to lie between 737 ms and 854 ms, corresponding to a predicted ambiguity disadvantage between 100 ms and 217 ms. The observed slowdown in the data was 128 ms. Thus the empirical mean of the ambiguous condition reading times was within the 95% confidence intervals for predicted reading times in all simulations. Consequently, the empirical mean did not significantly deviate from the predicted mean reading time (all ps>0.1, two-tailed). The SMCM with a first-terminating stopping-rule (which we have included only for illustration) predicted the reading time for ambiguous conditions to lie between 443 ms and 499 ms, corresponding to

an ambiguity *advantage* 194 ms and 138 ms.

7.3.3 Discussion

Our simulation results demonstrate that the empirical mean reading time in the ambiguous condition does not significantly differ from SMCM's predictions. This result held true under different distributional assumptions, which produced almost identical predictions and importantly, the reading time data from the ambiguous condition was not used in parameter-estimation. Our result does not constitute evidence against good-enough theory, because successive computation of two attachments is compatible with slowdowns of any magnitude in the ambiguous condition. The results of the simulation demonstrate however, that the SMCM makes more constrained predictions than a serial model, without additional assumptions and without free parameters, and can thus be more easily tested.

7.4 General Discussion

We have provided empirical evidence for the task-dependence of ambiguity-resolution. In addition, we have presented the SMCM, a new quantitative model that can explain the effect in terms of a non-deterministic model of parsing which assumes multiple channels and a stopping rule. This model can be seen as a refinement of the SDCF idea and is an improvement over the underspecification proposal because it makes precise assumptions about the timing of the attachment process, and because it makes precise predictions about the relationship between the attachment times in ambiguous and unambiguous conditions.

The SMCM has one free parameter, namely the kind of stopping rule used by the parser. We assume that the stopping rule selection depends on the type of questions participants are asked, with questions in the present experiment causing the parser to use an exhaustive stopping rule. We furthermore assume that all participants use the same stopping rule on all trials. Clearly, it is possible to imagine that the same task may prompt different people to make use of different stopping rules at different times, as they may differ in their perception of task demands. Furthermore, different participants may decide to start using an exhaustive stopping-rule at different points in the experiment, while others may use a first-terminating stopping-rule throughout. As a result, some tasks may result in the usage of an exhaustive stopping-rule on only a certain proportion of trials. Under such circumstances, the exact pre-

dictions of the SMCM will depend on this proportion, which would constitute an additional free parameter. Therefore, such a situation would not constitute a strong test of the SMCM unless we have an independent way of estimating this proportion, because this free parameter would enable the SMCM to explain a whole (restricted) range of reading times for ambiguous sentences with the right proportion parameter. However, the present data do not provide any indication that such a parameter may have to be used.

Importantly, although the SMCM is a parallel processing model with independent channels, it does not necessarily keep all the parses that have been computed. This means that it is not parallel in the sense that several parses are maintained (e.g., Lewis, 2000). We assume that the SMCM parser discards all structures except one, because of the low proportion of 'yes' responses to RC questions in our experiment. This is because a 'no' response to a question about a particular attachment indicates that an interpretation with this attachment was not in storage at the time the question was answered. Thus, a low proportion of 'yes' responses suggests a high number of trials on which the reading asked about was not in storage. We assume that this is because the other reading was selected on that trial. There are several possible reasons for the parser to retain only one parse. One possibility is that the number of stored structures is subject to task-demands as well. While our task did modify the parser's stopping behavior, a somewhat different task may be required to modify the parser's storage strategy. For instance, participants may have been more inclined to store multiple interpretations of a sentence if they had been provided with feedback as to the correctness of their answer. Another possibility is that the strategy of selecting only one parse when several are available is intrinsic to the parsing algorithm and non-modifiable by task demands. Structures may be discarded due to memory limitations, or because the parser's architecture forces it to be take a decision whenever several options are simultaneously available (e.g. Frazier, 1987).

Theories of sentence processing diverge on two main issues: (i) the number of analyses maintained by the parser at any given moment, and (ii) the algorithm underlying potential disambiguation. The timing of the processes which generate different parses is rarely addressed. Only two theories make explicit claims about the timing of these processes: The Garden-Path Model (Frazier, 1979, 1987), and the URM (van Gompel et al., 2000). Both theories assume that analyses are created simultaneously. For instance, according to the Garden-Path Model, a deterministic race is the mechanism underlying the parsing principle minimal attachment. When facing a

choice between two structures, the parser always adopts minimal structure because less time is required to build it. However, unlike the URM, the Garden-Path Theory assumes no stochasticity (non-determinism) in the structure-building process, which is why the minimal attachment structure *always* takes less time.

Interestingly, *cue-based parsing* may make similar predictions when faced with ambiguity. Lewis, Vasishth, and Van Dyke (2006) assumes that in order to establish a dependency between two co-dependent elements like subjects and verbs, the first-occurring co-dependent needs to be retrieved from memory. In post-nominal relative clauses, for example, RC attachment requires the parser to retrieve the noun phrase to which the RC attaches when it encounters the verb. This is because that noun phrase is an argument of the RC verb, and retrieval is assumed to precede dependency resolution. If only one reading needs to be computed, only one noun phrase needs to be retrieved. This operation may be carried out more quickly in the ambiguous condition than in an unambiguous condition because two candidate noun phrases are available in memory in the former case, as opposed to just one in the latter — and retrieval of any one of two noun phrases must finish faster on average due to a higher probability of the search finishing early.

Thus, while the exact nature of the operations engaged in a race is left open in the URM as well as the SMCM, an integration of cue-based parsing and SMCM could fill in this gap by assuming that at least a part of the computations performed simultaneously is due to search in content-addressable memory (McElree, 2000; Martin & McElree, 2009). We leave this question for future research, but fortunately, we are able to derive quantitative predictions even without the knowledge of the exact mechanism of these processes, because the reading times in the unambiguous conditions provide us with estimates of their durations.

7.5 Summary

In chapter three of this thesis it was shown through reanalysis of experimental data of Swets et al. (2008) and by simulation that they do not provide conclusive evidence for the assumption that the parser's handling of ambiguities depends on task demands. In chapters six and seven, evidence against strategic underspecification was presented. In the current chapter, experimental evidence was presented which supports the Swets et al. claim that different strategy resolution strategies can be employed during reading, and that task demands may determine which strategy is used. However, it does not support their claim that underspecification is responsible

for the ambiguity advantage. Given a sufficient amount of task difficulty, sentences with attachment ambiguities can be read more slowly than unambiguous sentences.

We then presented a new model, the stochastic multiple-channel model of ambiguity resolution (SMCM) as an extension of van Gompel et al.'s (2000) unrestricted race model. Finally, we demonstrated that in addition to being able to account for previous findings (Traxler et al., 1998; van Gompel et al., 2000, 2001, 2005) the SMCM makes more constrained predictions than a model which assumes sequential attachment. The SMCM is, to the best of our knowledge, the first quantitative model of task-dependent ambiguity resolution.

Chapter 8

Summary and Conclusion

In this thesis, two models of the parser's actions at choice-points in the sentence were presented and evaluated, theoretically, and with respect to experimental evidence. The *unrestricted race model* (URM) assumes that the parser non-deterministically resolves ambiguities whenever it encounters any. It assumes that the human comprehension system attempts to construct several permissible interpretations of a sentences at the same time, and adopts whichever interpretation is constructed fastest. The *strategic underspecification* model assumes that readers do not attempt to resolve ambiguities unless it is absolutely necessary. In other words, they *underspecify*. With these assumptions both models can explain the *ambiguity advantage*, which is the finding that ambiguous sentences are read faster than their unambiguous counterparts (e.g., Traxler et al., 1998). Chapters 1 and 3 introduce these two models, as well as the empirical evidence for them.

The primary aim of this thesis was to refine these competing theories of ambiguity resolution and to decide between them based on experimental evidence. The secondary aims were, firstly, to put the hypothesis of task-dependent ambiguity resolution to a further test and, secondly, to explore the hypothesis that the ambiguity advantage in reading is — at least in part — not caused by *faster* but by *more successful* processing.

Chapter 3 discusses Swets et al.'s (2008) argument against the URM. They presented experimental evidence showing that (i) the ambiguity advantage does not occur when the task encourages ambiguity resolution, and (ii) comprehension questions about a potentially ambiguous part of the sentence are answered more slowly when that part is ambiguous than when it is unambiguous. They argued that the strategic underspecification model predicts both effects. According to them, read-

ers underspecify very rarely when the task encourages ambiguity resolution because the parsing strategy is sensitive to the task. However, when they do underspecify, they require more time to answer comprehension questions because they need to disambiguate the sentence before answering a question. Because only ambiguous sentences can be underspecified, question-answering to such sentences is selectively slowed. Swets and colleagues further argued that the URM is incompatible with these findings because it is not susceptible to task demands, and therefore cannot explain the disappearance of the ambiguity advantage under task demands encouraging ambiguity resolution. Moreover, Swets et al. argue that the URM cannot explain the slow responses to questions about ambiguous sentences relative to unambiguous ones, because it assumes that sentence representations are always fully specified. These claims were evaluated in chapter three. Firstly, a reanalysis of Swets et al.'s question-response data showed that the statistical evidence for a slowdown in question-response times was inconclusive. As a result, their finding is not an argument against the URM. Concerning the disappearance of the ambiguity advantage, it was shown that a modified version of the URM is compatible with Swets and colleagues' findings. Specifically, the URM needs to assume that syntactic constituents are only integrated into the sentence after they have been processed to a sufficient degree. Importantly, this is an assumption that the strategic underspecification account needs to make as well. Simulations were presented, which showed that, under this assumption, the URM predicts a very small ambiguity advantage, which would not have been found in Swets et al.'s experiment due to lack of statistical power.

Chapter 4 discussed the precise assumptions of the strategic underspecification model. The result were two instantiations of the underspecification idea: *partial specification* and *non-specification*. The partial specification assumes that some information about possible meanings of an ambiguity is retained even when the parser underspecifies. As a result, it can use the underspecified representation to disambiguate it at a later point. The non-specification model assumes that no information about permissible meanings of the ambiguous part of the sentence is stored. The consequence is that the parser has no access to this information at a later point. Computational implementations of both models were able to account for Swets et al.'s data equally well, and so no conclusion as to the exact processes underlying underspecification can be made in the absence of further data. Interestingly, the parameter estimates of both models suggested that underspecification — to the extent it exists — might be deterministic. This is because according to the estimated parameters of both models, the probability of underspecification for each participant was either 1 or 0 in most cases.

Chapter 5 presented the results of a self-paced reading experiment concerning relative clause attachment in Turkish, for which the URM and the strategic underspecification model made diverging predictions. They do so because in Turkish, relative clauses precede the noun they modify, instead of following it as they do in English and in German. Thus, the typical order of a Turkish construction in which the relative clause could attach either to one of two nouns is *relative clause* $noun_1$ $noun_2$. The URM and the underspecification model make different predictions for ambiguous sentences because strategic underspecification aims to minimize its processing load, task demands allowing. The best strategy for minimizing processing effort in Turkish relative clauses is to delay disambiguation until encounter the second noun is encountered, and to underspecify ambiguous sentences, unless ambiguity resolution is required. As a result, strategic underspecification predicts an ambiguity advantage in Turkish. The URM, on the other hand, predicts that if the possibility of attaching a (fully processed) relative clause to a noun exists, the parser should do so immediately, regardless the task demands. Therefore, it predicts that disambiguation should happen on the first noun, whenever possible. Thus, there should be no ambiguity advantage. These results of the self-paced reading experiment, in which task demands did not encourage disambiguation, agree with the predictions of the URM, but not with those of strategic underspecification.

Chapter 6 provided a more direct test of the predictions of the two models of ambiguity resolution. The underspecification model assumes that ambiguous sentences are read faster because an entire processing step, i.e., disambiguation, is omitted. This has implications for the predicted completion time distributions of ambiguous and unambiguous sentences: the minimal amount of time required to process an ambiguous sentence should be shorter than the minimal amount of time required to process an unambiguous sentence. This prediction means that when the task is to classify sentences as grammatical or ungrammatical given a fixed amount of time, the earliest responses indicating successful processing of the sentence should occur earlier in ambiguous sentences than in unambiguous sentences. The URM, on the other hand predicts no such differences. This is because processing an ambiguous sentence does not involve qualitatively different processing. According to the URM, ambiguities are compatible with more interpretations, and therefore have a higher chance of being processed quickly, because it is more likely that one of several interpretations being constructed in parallel is completed quickly than that one single interpretation will be constructed quickly. However, the minimal amount of time to do so, over all experimental trials, should not differ. This prediction was tested in a response-signal paradigm experiment with German relative clauses. Counter the

predictions of the strategic underspecification model, the minimal amount of time required to accurately assess the grammaticality of a sentence did not differ between ambiguous and unambiguous sentences. The results were in agreement with the predictions of the URM, but not with those of underspecification. In addition, the results showed that participants were more accurate at identifying ambiguous sentences as acceptable than their unambiguous counterparts. This means that participants are more successful at assigning valid syntactic representations to ambiguous sentences. This finding is predicted by the URM, but not by the strategic underspecification model.

The final chapter 7 asks the question: can the human sentence comprehension system adapt to task demands with respect to the choices it makes? As discussed in chapter three, Swets and colleagues do not provide conclusive evidence that end. For this reason, a third experiment is presented, which involved a German relative clause attachment ambiguity, and task demands which encourage readers to construct and retain both interpretations of ambiguous structures. The experiment shows an ambiguity *disadvantage*, which suggests that readers do indeed compute both interpretations of ambiguous sentences, because constructing both meanings should require more time than constructing only one. Because a model of ambiguity resolution should be able to account for both effects, the *stochastic multiple-channel model* (SMCM) of ambiguity resolution is proposed, which is susceptible to task demands in a limited way. Like the URM, the SMCM assumes that the parser attempts to compute several interpretations in parallel. However, unlike the URM, it is not limited in the number of analyses it can generate. Exactly how many analyses are computed, is determined by a *stopping rule*, which is chosen based on the task. If the stopping rule is *first-terminating*, the parser makes the same prediction as the URM: an ambiguity advantage. If the stopping rule is *exhaustive*, it predicts an ambiguity *disadvantage*. The data from the reading time experiment is used to test the quantitative predictions of the SMCM, and no significant difference between the model predictions and the reading times in the ambiguous condition is found.

The present thesis first argued that the argument in favor of task-dependent strategic underspecification put forward by (Swets et al., 2008) is not compelling. Moreover, the precise assumptions of the underspecification model concerning the structure of underspecified representations are not entirely clear. Therefore, the notion of what it means to underspecify was sharpened by specifying the precise structure of two such models. Subsequently, empirical evidence against strategic underspecification and in favor of the URM was presented. This evidence suggests that readers can com-

pute several interpretations of ambiguous sentences at the same time. Furthermore, the response-signal paradigm experiment shows that human sentence comprehension is fallible, and that it may be so to different degrees in different experimental conditions. Thus models of sentence comprehension need to take into account the probability of successful computation of structure in addition to the time that such computations require. Finally, the evidence for the influence of task demands on disambiguation strategies suggests that task effects may play a major role in sentence comprehension research and need to be accounted for.

Appendix A

Models 2 and 3: Details

In order to compute the likelihood of the data, given a set of parameters, which is necessary for maximum-likelihood estimation (MLE), we assumed that all reading times and reaction times are distributed according to a gamma distribution. We assumed assumed a common scale parameter of of each participant, because the distribution function of a sum of two random variables (RVs) can be obtained by simply adding the shape parameters under this assumption. Obtaining the distribution of a sum of two RVs is necessary in order to model the assumption that N1 and N2 attachment should require the same amount of time, irrespective of whether they occur during reading or during question-answering.

A.1 Partial specification

Figure A.1 provides an overview of the kinds of trials assumed by the partial specification model. All reading times and reaction times were assumed to follow a gamma distribution with a common scale parameter θ. Thus, differences in process duration were modeled as differences in the shape parameter κ. Table A.1 shows κs for reading and reaction times on the different types of trials illustrated in figure A.1. α_0 is the shape parameter for the average reading time for one word. α_1 corresponds to a difference in κ between underspecification trials and N2 attachment trials. α_2 corresponds to a difference in κ between N2 attachment and N1 attachment trials. β_0 is the shape parameter for a regular informed response on trials where attachment was carried out during reading. β_1 corresponds to a difference in κ between informed responses and guesses. Finally, w_n corresponds to the number of words making up the critical region, and is a property of the sentence and not a param-

Table A.1: Partial specification: Parameter assumptions for different types of trials.

	reading time κ	response time κ	$P(response = N1)$	P(occurrence in ambiguous)
Type 1A	$n_w\alpha_0 + \alpha_1 + \alpha_2$	β_0	1	$(1-p_u)(1-p_{N2})(1-p_{err})$
Type 1B	$n_w\alpha_0 + \alpha_1$	β_0	0	$(1-p_u)(1-p_{N2})p_{err}$
Type 2A	$n_w\alpha_0 + \alpha_1 + \alpha_2$	$\beta_0 + \beta_1$	0.5	$(1-p_u)p_{N2}(1-p_{err})$
Type 2B	$n_w\alpha_0 + \alpha_1$	$\beta_0 + \beta_1$	0.5	$(1-p_u)p_{N2}p_{err}$
Type 3A	$n_w\alpha_0$	$\beta_0 + \alpha_1 + \alpha_2$	1	$p_u(1-p_{N2})(1-p_{err})$
Type 3B	$n_w\alpha_0$	$\beta_0 + \alpha_1$	0	$p_u p_{N2}(1-p_{err})$
Type 3C	$n_w\alpha_0$	$\beta_0 + \beta_1$	0.5	$p_u(1-p_{err})$

eter to be estimated. We assumed that all α_i and $\beta_i \geq 20/\theta$, to ensure that the minimum difference between two processes of which one is supposed to be slower is $20\,ms$ or more. We assumed three further free parameters: p_{err}, p_u, and p_{N2}. The probability of not being able to retrieve the sentence during question answering was p_{err}. Thus, trials of type 2A and 2B occurred with a probability of p_{err} in the N1 and N2 attachment conditions, respectively. The probability of underspecifying attachment in the ambiguous condition was p_u, and so trial types 3A and 3B occurred with a probability of p_u in the ambiguous condition. Finally, p_{N2} quantified the preference for attaching to N2. Thus, the probabilities of trial types 1A and 1B in the ambiguous condition were $(1-p_u)(1-p_{N2})$ and $(1-p_u)p_{N2}$, respectively. Accordingly, the probabilities of trial types 3A and 3B in the ambiguous condition were $p_u(1-p_{N2})$ and $p_u p_{N2}$, respectively.

Figure A.1: Different kinds of trials in the three attachment conditions according to the partial specification model.

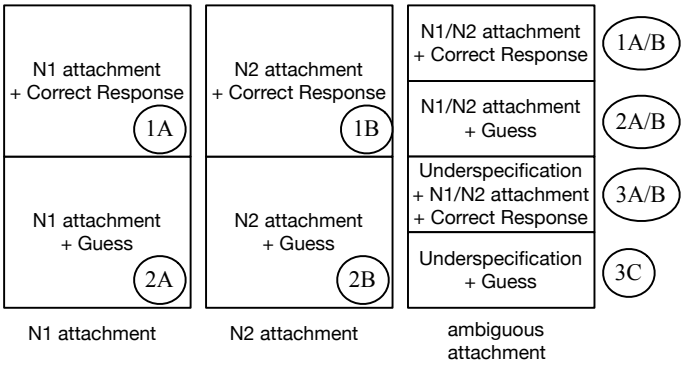


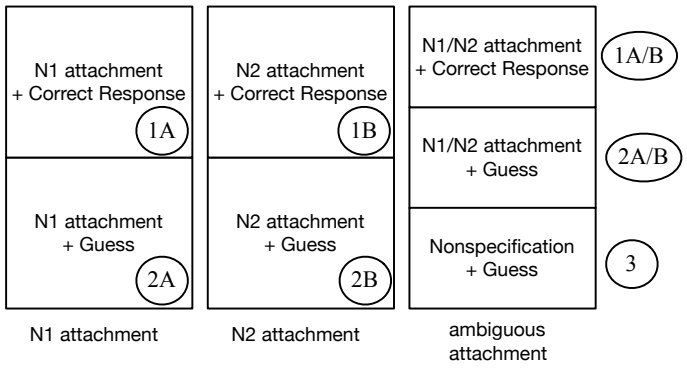
Table A.2: Non-specification: Parameter assumptions for different types of trials.

	reading time κ	response time κ	$P(response = N1)$	P(occurrence in ambiguous)
Type 1A	$n_w\alpha_0 + \alpha_1 + \alpha_2$	β_0	1	$(1 - p_u)(1 - p_{N2})(1 - p_{err})$
Type 1B	$n_w\alpha_0 + \alpha_1$	β_0	0	$(1 - p_u)(1 - p_{N2})p_{err}$
Type 2A	$n_w\alpha_0 + \alpha_1 + \alpha_2$	$\beta_0 + \beta_1$	0.5	$(1 - p_u)p_{N2}(1 - p_{err})$
Type 2B	$n_w\alpha_0 + \alpha_1$	$\beta_0 + \beta_1$	0.5	$(1 - p_u)p_{N2}p_{err}$
Type 3	$n_w\alpha_0$	$\beta_0 + \beta_1$	0.5	p_u

A.2 Non-specification

Figure A.2 provides an overview of the kinds of trials assumed by the partial specification model. Here to, all reading times and reaction times were assumed to follow a gamma distribution with a common scale parameter θ. Table A.2 shows κs for reading and reaction times on the different types of trials illustrated in figure A.2. Distributions and response probabilities for trial types 1A, 1B, 2A and 2B are exactly the same as in the partial specification model. The interpretation of the α_i and β_i parameters as well as the constraints on them were the same, as well. We also made use of the parameters p_{err}, p_u, and p_{N2} in this model. The crucial difference, however, is that the non-specification model assumes that attachment can only occur during reading. Thus, all underspecification trials, occurring with a probability of p_u, are followed by guessing in the question-answering phase.

Figure A.2: Different kinds of trials in the three attachment conditions according to the non-specification model.



References

Albers, W., & Kallenberg, W. (1994). A simple approximation to the bivariate normal distribution with large correlation coefficient. *Journal of multivariate analysis*.

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge, MA: Cambridge University Press.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013, April). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4.*

Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (chap. 9). Wiley, New York.

Bolker, D. B., Maechler, M., Bates, D., & Walker, S. (2013). *lme4: Linear mixed-effects models using S4 classes (R package version).*

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior research methods, instruments, & computers*, *35*(1), 11–21.

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.

Colonius, H., & Vorberg, D. (1994). Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology*, *38*(1), 35–58. doi: M10.1006/jmps.1994.1002

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.

Dosher, B. A. (1979). Empirical approaches to information processing: Speed-accuracy tradeoff functions or reaction time—A reply. *Acta Psychologica*, *43*, 347–359.

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.

Foraker, S., & McElree, B. (2007). The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, *56*(3), 357–383. doi: M10.1016/j.jml.2006.07.004

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies* (Doctoral dissertation). University of Connecticut.

Frazier, L. (1987). Sentence Processing: A Tutorial Review. In *Attention and performance xii: The psychology of reading* (pp. 559–586). London: Erlbaum.

Frazier, L., & Clifton, C. J. (1996). *Construal*. Cambridge, MA: MIT Press.

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

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of naacl 2001* (Vol. 2, pp. 159–166). Morristown, NJ, USA: Association for Computational Linguistics.

Hemforth, B., Konieczny, L., & Scheepers, C. (2000). Syntactic attachment and anaphor resolution: Two sides of relative clause attachment. In M. Crocker, M. Pickering, & C. Clifton (Eds.), *Architectures and mechanisms for language processing* (pp. 259–282). Cambridge University Press.

Hesterberg, T., Moore, D. S., Clipson, A., & Epstein, R. (2005). Bootstrap methods and permutation tests. In *Introduction to the practice of statistics* (5th ed., pp. 1–70). New York: W.H. Freeman and Company.

Johnson, S. G. (2013). *The NLopt nonlinear-optimization package.* Retrieved from `Mhttp://ab-initio.mit.edu/nlopt`

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2002). Perspective Effects on Online Text Processing. *Discourse Processes*, *33*(2), 159–173.

Lacouture, Y., & Cousineau, D. (2008). How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 35–45.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi: M10.1016/j.cognition.2007.05.006

Lewis, R. L. (2000). Falsifying serial and parallel parsing models: Empirical conundrums and an overlooked paradigm. *Journal of Psycholinguistic Research*,

29(2), 241–248.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, *51*(5), 341–352.

Liu, C. C., & Smith, P. L. (2009). Comparing time-accuracy curves: beyond goodness-of-fit measures. *Psychonomic bulletin & review*, *16*(1), 190–203. doi: M10.3758/PBR.16.1.190

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory (2nd edition)*. Mahwah, NJ: Erlbaum.

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*(3), 879–906. doi: M10.1016/j.jml.2007.06.010

Martin, A. E., & McElree, B. (2009). Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1231.

Martin, A. E., & McElree, B. (2011). Direct-access retrieval during sentence comprehension: Evidence from Sluicing. *Journal of memory and language*, *64*(4), 327–343. doi: M10.1016/j.jml.2010.12.006

McConkie, G. W., Rayner, K., & Wilson, S. J. (1973). Experimental manipulation of reading strategies. *Journal of Educational Psychology*, *66*(1), 1–8.

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language*(32), 536–571.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*(2), 247–279.

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

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100.

Myung, I. J., & Pitt, M. (2009). Optimal experimental design for model discrimination. *Psychological review*, *116*(3), 499–518. doi: M10.1037/a0016104

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*.

Pachella, R. G. (1974). The interpretation of reaction time in information processing

research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition.* Hillsdale, N.J.: Erlbaum.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus.* New York: Springer.

Pitt, M., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421–425.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raab, D. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*, 574–590.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology.* Cambridge, MA: Blackwell.

Rowan, T. (1990). *Functional stability analysis of numerical algorithms* (Unpublished doctoral dissertation).

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & von der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*(4), 583–639.

Stan Development Team. (2014). *Stan: A C++ Library for Probability and Sampling, Version 2.2.* Retrieved from `Mhttp://mc-stan.org/`

Stephan, K. E., & Penny, W. D. (2006). Dynamic causal models and Bayesian selection. *Statistical parametric mapping: the analysis of functional brain images.*, 577–585.

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.

Townsend, J. T., & Ashby, G. F. (1983). *Stochastic Modeling of Elementary Psychological Processes.* Cambridge, MA: Cambridge University Press.

Townsend, J. T., & Colonius, H. (1997). Parallel Processing Response Times and Experimental Determination of the Stopping Rule. *Journal of mathematical psychology*, *41*(4), 392–397.

Townsend, J. T., & Honey, C. J. (2007). Consequences of base time for redundant signals experiments. *Journal of Mathematical Psychology*(51), 1–24.

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.

Ulrich, R., & Miller, J. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, *37*, 513–525.

Van Dyke, J. a., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263. doi: M10.1016/j.jml.2011.05.002

Van Zandt, T. (2002). Analysis of response time distributions. *Psychonomic Bulletin {&} Review*, 1–99.

van Gompel, R. P. G., & Pickering, M. J. (2006). Syntactic parsing. In M. J. Traxler & M. A. Gernsbacher (Eds.), *The oxford handbook of psycholinguistics* (2nd ed., pp. 289–307). London: Academic Press.

van Gompel, R. P. G., 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.

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.

van Gompel, R. P. G., 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: M10 .1006/jmla.2001.2773

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S-PLUS.* New York: Springer.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779–804.

Wagenmakers, E.-J., & Brown, S. (2007). On the Linear Relation Between the Mean and the Standard Deviation of a Response Time Distribution. *Psychological Review*, *114*(3).

Wehrens, R., Putter, H., & Buydens, L. M. C. (2000). The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *54*(1), 35–52.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85.

Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. (1980). Priming and retrieval from short-term memory: A speed accuracy trade-off analysis. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 387–404.

Wotschack, C. (2009). *Eye Movements in Reading Strategies* (Doctoral dissertation). Unversität Potsdam.