Toward an Integrated Model of Sentence Processing in Reading

Felix Engelmann

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

University of Potsdam
2016

Supervisors:

Prof. Dr. Shravan Vasishth
Prof. Dr. Reinhold Kliegl

Submitted: November 30, 2015
Defended: June 1, 2016

Reviewers:

Prof. Dr. Shravan Vasishth, University of Potsdam, Germany
Prof. Dr. Hedderik van Rijn, University of Groningen, The Netherlands

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-100864
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-100864

Abstract

In experiments investigating sentence processing, eye movement measures such as fixation durations and regression proportions while reading are commonly used to draw conclusions about processing difficulties. However, these measures are the result of an interaction of multiple cognitive levels and processing strategies and thus are only indirect indicators of processing difficulty. In order to properly interpret an eye movement response, one has to understand the underlying principles of adaptive processing such as trade-off mechanisms between reading speed and depth of comprehension that interact with task demands and individual differences. Therefore, it is necessary to establish explicit models of the respective mechanisms as well as their causal relationship with observable behavior. There are models of lexical processing and eye movement control on the one side and models on sentence parsing and memory processes on the other. However, no model so far combines both sides with explicitly defined linking assumptions.

In this thesis, a model is developed that integrates oculomotor control with a parsing mechanism and a theory of cue-based memory retrieval. On the basis previous empirical findings and independently motivated principles, adaptive, resource-preserving mechanisms of underspecification are proposed both on the level of memory access and on the level of syntactic parsing. The thesis first investigates the model of cue-based retrieval in sentence comprehension of Lewis and Vasishth (2005) with a comprehensive literature review and computational modeling of retrieval interference in dependency processing. The results reveal a great variability in the data that is not explained by the theory. Therefore, two principles, *distractor prominence* and *cue confusion*, are proposed as an extension to the theory, thus providing a more adequate description of systematic variance in empirical results as a consequence of experimental design, linguistic environment, and individual differences. In the remainder of the thesis, four interfaces between parsing and eye movement control are defined: *Time Out*, *Reanalysis*, *Underspecification*, and *Subvocalization*. By comparing computationally derived predictions with experimental results from the literature, it is investigated to what extent these four interfaces constitute an appropriate elementary set of assumptions for explaining specific eye movement patterns during sentence processing. Through simulations, it is shown how this system of in itself simple assumptions results in predictions of complex, adaptive behavior.

In conclusion, it is argued that, on all levels, the sentence comprehension mechanism seeks a balance between necessary processing effort and reading speed on the basis of experience, task demands, and resource limitations. Theories of linguistic processing therefore need to be explicitly defined and implemented, in particular with respect to linking assumptions between observable behavior and underlying cognitive processes. The comprehensive model developed here integrates multiple levels of sentence processing that hitherto have only been studied in isolation. The model is made publicly available as an expandable framework for future studies of the interactions between parsing, memory access, and eye movement control.

Acknowledgments

I owe undying gratitude to Shravan Vasishth, who is the most caring and encouraging teacher and adviser I can imagine. He always provided the perfect mixture of demand and patience for my needs, while being a reliable source of enthusiasm and confidence whenever I was in need for it. Thank you for putting so much work into my work and for making me a researcher.

I want to thank Reinhold Kliegl and Ralf Engbert for their advice as well as a fruitful collaboration leading to my first journal publication that kick-started this thesis. Furthermore, I am grateful to Hedderik van Rijn for his numerous inspirational comments on my work and for being my ACT-R modeling mentor. I also thank Dario Salvucci for providing his code and kind support for the replication of his EMMA simulations.

I am grateful to the Potsdam Graduate School, which covered my travel costs for several conferences and granted me a five months PhD Completion Scholarship.

I especially want to thank Lena Jäger, not only for her thorough proof-reading of parts of the thesis, but also for all these creative times that we were developing ideas together that culminated into two joint papers and one chapter of this thesis. In addition, I cannot overstate the role of my friends and colleagues Titus von der Malsburg, Bruno Nicenboim, Pavel Logaĉev, Paul Metzner, Umesh Patil, Rukshin Shaher, and Samar Husain as sources and challengers of ideas, as collaborators and advisers for the whole time of my PhD.

Finally, I would like to say thank you to my parents for their infinite and unconditioned support and patience without ever completely understanding what I was actually doing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

At the 2015 CUNY sentence processing conference in Los Angeles, after a talk, a discussion evolved about whether one can say that human communication is “easy” or that it is “hard”. This on first sight very simple question turned out to be surprisingly controversial. Communication seems easy because humans do it every day, all the time without thinking much about it. But seen from the perspective of a researcher that tries to sort out all influential factors and their interactions constituting successful communication, it is a very hard task. One could probably settle on the assertion that communication is *complex*, but humans are highly trained experts with many years of practice, hence, it at least *looks* easy. But note that communication is many times not successful and never perfect. It, nevertheless, mostly *works* because humans have developed efficient strategies to repair or accommodate failed or insufficient communication. This is possible because communication always happens in an informative context, and it is necessary because communication takes place under — internal or external — time constraints. External time constraints are imposed on a listener who has to keep up with the speaker’s pace. But also without external time constraints, e.g., in reading, a comprehender tries to maintain a certain speed and avoid interruptions. That is, the ability to communicate has developed as an economically efficient process that tries to be at the same time as precise as necessary and as fast as possible. In order to understand how communication works, we as researchers have to account for the factors that make it “hard” *and* uncover the cognitive strategies that make it look “easy”. The observational tools at hand are, however, limited: Experimental methods, such as response time measuring, eye tracking, or the recording of electrophysiological brain responses, measure observable reactions to unobservable cognitive processing which might be an interaction of functional constraints and strategic adaptation. The challenge is to disentangle one from the other. It is therefore essential to develop explicit models of the way high-level cognitive processing is linked to observable behavior through low-level processes and motor control.

In sentence processing in reading, which is the topic of this thesis, it is undeniable that the relation between parsing and eye movements is mediated by, e.g., word recognition processes, the allocation of attention, and the oculomotor control system. To date, the interpretation of eye movement data is built on rather vague assumptions about the interaction of linguistically relevant cognition with these low-level functions.

The usual way to analyze eye-tracking data is to compute a range of "early" and "late" fixation measures as to uncover the timing of parsing processes, thereby often assuming a fairly direct relation between between a point of difficulty and the eye movement response. A slow-down in a measure or a backward movement (regression) is usually interpreted as signaling difficulty. Early events are expected to have effects in early measures such as first fixation and first pass reading time (also called gaze duration), and later processing stages are supposed to affect second pass measures such as the rereading of a word or phrase (Clifton, Staub, & Rayner, 2007). In one of the first comprehensive models of reading, Just and Carpenter (1980) proposed an immediacy between processing and eye movement control, termed the "eye-mind assumption":

> The eye-mind assumption posits that there is no appreciable lag between what is being fixated and what is being processed (p. 331).

Research has shown, however, that neither the timing nor the quality of the response are very tightly linked to parsing processes. First of all, low-level oculomotor constraints alone may cause late effects to spill over into early measures of adjacent regions (Rayner & Duffy, 1986), and parafoveal preview of upcoming words can influence fixation latencies on the currently fixated word (Rayner, 1998). Regarding post-lexical processing, Boston, Hale, Vasishth, and Kliegl (2011) have shown that metrics of parsing difficulty such as surprisal (Hale, 2011; Levy, 2008) and memory retrieval (Lewis & Vasishth, 2005) are correlated with both early and late measures. Furthermore, as stated by the theories of good-enough processing (Ferreira, Ferraro, & Bailey, 2002; Sanford & Sturt, 2002) and construal (Carreiras & Clifton, 1993; Frazier & Clifton, 1997), processes and decisions are sometimes abandoned or deferred, such that effects of difficulty can appear several words downstream, be completely absent, or even appear as facilitation (von der Malsburg & Vasishth, 2013; Swets, Desmet, Clifton, & Ferreira, 2008; Traxler, 2007; Traxler, Pickering, & Clifton, 1998; Van Gompel, Pickering, & Traxler, 2001). When regressions are observed, their functions and their targets are often not directly connected to immediate parsing difficulty (von der Malsburg & Vasishth, 2011; Meseguer, Carreiras, & Clifton, 2002; Mitchell, Shen, Green, & Hodgson, 2008). And finally, the reactions to, for example, ambiguous phrases vary with individual differences and task demands (von der Malsburg & Vasishth, 2013; Swets et al., 2008).

The most promising way toward a better understanding of the complexity described above is to build explicit linking-assumptions into falsifiable computational models. The apparently tight link between word-level factors and eye movements has enabled the development of process models that define the interactions between oculomotor constraints and lexical and visual word information. On the word-level, fixation durations and locations are highly predictable by the visual and lexical properties of the material. Factors like lexical frequency, word length, and lexical predictability show immediate effects on fixations (Kliegl, Grabner, Rolfs, & Engbert, 2004; Rayner, 1998). The two most advanced models in this area are E-Z Reader (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle et al., 2009) and SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005). Both specify motor planning of saccades and accurately predict individual fixation durations, saccade landing sites, as well as phenomena like parafoveal preview. Interestingly, however, the two models rely on different assumptions about attention allocation: E-Z Reader shifts attention serially, whereas SWIFT uses an attentional gradient to process several words in parallel. Although models of this category are very advanced, they still illuminate only part of the picture. Validation is usually done on eye-tracking corpora that consist of particularly simple sentences, eliminating the influence of 'higher-level' processes concerned with the structured analysis of inter-word dependencies (note, however, Reichle et al., 2009's approach toward the integration of high-level influences described below).

Models that deal with post-lexical processing, on the other hand, neglect the 'low-level' functions that build the connection to observable behavior. This class of models defines rather abstract metrics of difficulty to specific regions in a sentence depending on structural attachment strategies (garden-path theory, Frazier & Fodor, 1978), syntactic expectation (Hale, 2001; Levy, 2008), constraint-satisfaction (MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey & Tanenhaus, 1998), structural frequency and regularity (Elman, 1990; Elman, Hare, & McRae, 2004; MacDonald & Christiansen, 2002), or memory constraints (Gibson, 2000; Just & Carpenter, 1992; Lewis & Vasishth, 2005; Van Dyke & Lewis, 2003). What these models are missing is an operationalized definition of difficulty. At least, the cue-based retrieval model suggested by Lewis and Vasishth (2005) and Van Dyke and Lewis (2003) quantifies timing differences in cognitive processes such as memory retrieval and the application of parsing rules (also see Boston et al., 2011). However, the mechanistic link between retrieval time and fixation time remains unaddressed in the model. The reasons why there are no explicit linking hypotheses to connect parsing and eye movements are twofold: On the one hand, the complexity of the eye-parser relation might have been underestimated. On the other hand, developing a model that accommodates the evidence about the mentioned interactions is an extremely challenging task.

The consequence is that, not only for models but also in the design of experiments, researchers either restrict themselves to studying low-level lexical processes ignoring higher-level influences or, when studying higher-level processing, neglect the influence of oculomotoric factors on the outcome of their experiments.

The only model to my knowledge that to some extent integrates both sentence-level information and eye movement control is Bicknell and Levy (2010). Their model predicts the timing and location of saccades and regressions as the result of a rational strategy guided by Bayesian inference on the sentence level.

The position advocated in this thesis is that, on all levels, the sentence comprehension mechanism seeks a balance between necessary processing effort and reading speed, using experience-based heuristics and taking into account task demands and individual capacity limitations. As a step toward a better understanding of these adaptive strategies, this thesis presents (a) an advancement to the standardly used theory of cue-based retrieval in sentence processing (Lewis & Vasishth, 2005) with respect to memory interference, including a mechanism of strategic underspecification in memory retrieval, (b) tests an explicit model that links parsing difficulty to eye movement control, and proposes (c) a dynamic interaction between parsing processes, individual differences, task demands, and a largely autonomous eye movement control, which naturally predicts the above mentioned balance between speed and accuracy. The process of cue-based retrieval is in itself an example of a cognitive mechanism adapted to an efficiency-guided trade-off between speed and accuracy as it sacrifices any information on serial order in favor of a rapid associative memory access. Most of the work presented here builds on the well-established computational model of cue-based retrieval parsing by Lewis and Vasishth (2005), implemented in the cognitive architecture ACT-R (Anderson et al., 2004; Anderson & Lebiere, 1998). The model of Lewis and Vasishth is extended with the prediction of individual fixations by implementing and testing an interface between parsing and eye movement control inspired by Mitchell et al. (2008) and Reichle et al. (2009). The proposed model should help to understand how high-level cognitive processes in reading interact with low-level motor control of eye movements. It is the first model to integrate both aspects of reading in one system.

The remainder of the introduction presents a brief overview of previous research on eye movements in sentence processing with regard to three topics which are to be addressed in the following chapters: Memory retrieval, the eye-parser interface, and strategic underspecificaton. The final section then summarizes the contents of the thesis.

1.1 Memory Retrieval

In sentence processing, structurally integrating a word means to build dependencies, e.g., between verbs and their arguments, or anaphors and their antecedents. For that, constituents have to be stored in and retrieved from working memory — a process that is affected by the distance between the dependents and by the amount of interference from similar memory items (Bartek, Lewis, Vasishth, & Smith, 2011; Gibson, 2000; Grodner & Gibson, 2005; Just & Carpenter, 1980, 1992).

The question of when and how interference from syntactically unlicensed items arises during the formation of dependencies is subject to an ongoing debate in the psycholinguistic community with a variability in the results that is hard to interpret. Some studies have found that an item that is not in the grammatically correct position can nevertheless interfere with the successful building of a dependency between two other items. In reflexive pronoun anaphora, Sturt (2003) found that reading times were influenced by gender (mis-)match with both the grammatically correct antecedent and a grammatically incorrect *distractor*. Effects of the distractor were only found in later measures. Other studies, however, have found an effect also in early measures (Jäger, Engelmann, & Vasishth, 2015; Patil, Vasishth, & Lewis, 2012). Cunnings and Felser (2013) found an effect that was modulated by working memory capacity. Again others (Dillon, Mishler, Sloggett, & Phillips, 2013; Wagers et al., 2009) have not found an effect in reflexives at all. As a consequence, there is no definite answer to the question when the retrieval of an antecedent is based purely on grammatical constraints and when features like gender are used to access items in memory. In addition, the directions of some of the observed effects are not consistent with the theory of cue-based retrieval, which is commonly used as an explanation for interference effects. Interestingly, in subject-verb dependencies, interference effects are more reliable but in many cases not compatible with the predictions of cue-based retrieval either.

In general, contradictory results in the literature can of course have a variety of reasons, but one in many cases very likely explanation is a lack of power in experimental studies, leading to Type I and Type II errors. Other reasons might be unaccounted effects of experimental design. Acknowledging this problem strengthens the importance of developing comprehensive models that make quantitative predictions, which the empirical findings can be evaluated against. Generating clear predictions with a computational model that takes into account experimental and individual differences can show what variations in the data are expected under the theory, especially when the interactions of influencing factors are too complex for verbally specified predictions. With computational models, it can be investigated in which way (a) unanticipated consequences of the cognitive processing mechanism and (b) the eye-parser interaction influence observable behavior.

This thesis contributes to both (a) the specification of the memory mechanisms of sentence parsing and (b) the development of an explicit eye-parser interface with the aim to provide a foundational step toward a fully integrated model of human sentence processing. The heterogeneous literature on memory interference in dependency resolution will be taken as a basis to advance the theory of cue-based retrieval in Chapter 3.

1.2 The Eye-Parser Interface

A particularly interesting case for the study of the eye-parser connection are regressive eye movements between words. Inter-word regressions are interruptions of the reading process that often indicate severe parsing difficulties or even a breakdown: Regressions can be induced by stimuli containing distant or embedded dependencies, ambiguous structures, or unexpected continuations. Consequently, there has been a growing interest in identifying spatio-temporal distributions of short- and long-range regressions with the objective to uncover constraints on memory operations and strategies of repair (Frazier & Rayner, 1982; von der Malsburg & Vasishth, 2011, 2013; Meseguer et al., 2002; Mitchell et al., 2008; Van Dyke & Lewis, 2003; Weger & Inhoff, 2007).

Results of an early study on temporal ambiguities by Frazier and Rayner (1982) seemed to imply a close relationship of the launch site and target of regressions with the supposed actions of the sentence processor. Frazier and Rayner studied sentences that contained a temporal ambiguity as in Example (1):

(1) Since Jay jogs a mile seems like a short distance to him.

This is a classical example of the so-called subject/object ambiguity: The parser, following the heuristic principle of *late closure* (garden-path theory, Frazier & Fodor, 1978), initially interprets the noun phrase *a mile* as the object of the subordinate clause (*jogs*), realizing at *seems* that *a mile* is actually the subject of the main clause. The parser is *garden-pathed* and tries to repair the wrongly built structure. Frazier and Rayner found evidence for overt *selective reanalysis*, sending the eyes back from the disambiguating region *seems* directly to the ambiguity that caused the error (*a mile*). However, later studies showed that the coupling of parsing and eye movements is not as tight as assumed. Regressions were found to be highly influenced by spatial properties of the text (Mitchell et al., 2008). As a contrasting proposal to Frazier and Rayner's "Selective Reanalysis", Mitchell et al. (2008) proposed the "Time Out hypothesis", which states that a regression is triggered at points of difficulty without any linguistically informed target, just for the purpose of interrupting the forward movement and giving the parser time to catch up with processing. In order to account for their data, they concluded that "regression sequences triggered by syntactic disambiguation are placed under relatively loosely-coupled control with the guidance of successive fixations being shared between both linguistic and non-linguistic mechanisms" (p. 285). Meseguer et al. (2002) reported further evidence for selective reanalysis in Spanish sentences, but their conclusions were weakened by a reexamination of their data with a novel method of scanpath analysis by von der Malsburg and Vasishth (2011). Von der Malsburg and Vasishth found two alternative patterns besides selective reanalysis: Rereading of the whole sentence and short one-word regressions, the latter of which is consistent with Mitchell et al.'s Time Out hypothesis. Also other researchers showed a mix of linguistic and spatial guidance for regressions related to reanalysis: Inhoff and Weger (2005) and Weger and Inhoff (2007) found that regressions that eventually reach the critical target use several steps, the first of which seems purely spatially guided, while the following corrective regressions show linguistic influence.

In a recent version of E-Z Reader, Reichle et al. (2009) provided a first proposal of how to connect eye movement control with high-level processes. They introduced a "placeholder" for a post-lexical integration stage that, with predefined probabilities, triggered returns to the difficult word or to an earlier region. Two types of integration failure were proposed: (a) "Slow integration failure" initiated a regression whenever the integration of word n is not completed before word n+1 has been lexically processed (similar to the "Time Out" hypothesis proposed by Mitchell et al., 2008); (b) "Rapid integration failure" initiated a regression immediately in the case that integration fails for some reason. Note, that the model did not incorporate any parsing theory that would predict *when* integration would fail or be delayed or *what* parsing processes lead to this. Reichle and colleagues focused solely on the question of *how* the eye movement system interacts with potential interruptions of the parsing process. Due to the lack of linguistic theory, the model also has no way of predicting regressions into earlier parts of the sentence than word n-1. Nevertheless, this is a promising step toward a fully specified model of parsing and eye movements and serves as inspiration for the model presented in Chapter 4.

1.3 Strategic Underspecification

In an eye-tracking experiment using a similar design as Meseguer et al. (2002), von der Malsburg and Vasishth (2013) found evidence for the “Good-enough” proposal, which suggests that the parser does not always build a complete structural representation of the sentence, but instead relies on effort-saving heuristics (Ferreira et al., 2002). Von der Malsburg and Vasishth report *faster* reading times in the ambiguous region compared to unambiguous sentences, suggesting that attachment processes might be suspended or abandoned in cases where the correct attachment site is not clear, leaving the attachment *underspecified.* An *ambiguity advantage* of this kind had been found before by other researchers in the potentially disambiguating region of ambiguous relative clauses (Swets et al., 2008; Traxler, 2007; Traxler et al., 1998; Van Gompel et al., 2001). In addition, von der Malsburg and Vasishth (2013) found that underspecification was modulated by individual differences in working memory capacity: Readers with higher capacity seemed to complete the attachment process more often, resulting in more reanalysis attempts at the disambiguating region, indicated by higher regression rates. It seems, furthermore, that underspecification is selectively applied in adjunct and relative clause attachments (von der Malsburg & Vasishth, 2013; Swets et al., 2008; Traxler et al., 1998) but not in ambiguous attachments that affect the relation between a verb and its arguments (e.g., Kemper, Crow, & Kemtes, 2004) and are therefore critical for comprehension. This selective strategy is consistent with the distinction between “primary” and “non-primary” attachment relations in *construal* theory (Frazier & Clifton, 1997). Further support for selective strategies comes from the finding that underspecification is influenced by task demands: Swets et al. (2008) found an ambiguity advantage only when using superficial comprehension questions but not in cases when the resolution of the ambiguous attachment was critical for answering the comprehension questions. The fact that the above mentioned studies found attachment decisions to be modulated by task demands and individual differences leads to the conclusion that readers adapt the depth of processing to their current goals and constraints in order to preserve an uninterrupted reading flow whenever possible. This will be addressed in a model in Chapter 5.

1.4 Structure of the Thesis

The thesis computationally investigates the mechanisms of sentence comprehension in reading on two levels. The first part subjects the classical cue-based retrieval theory of Lewis and Vasishth (2005) to a thorough assessment with respect to a large range of experimental evidence and proposes an extension to cover adaptive processing. In the second part of the thesis, four interfaces between parsing and eye movement control are established as an elementary set to predict reading behavior and strategic underspecification.

First, Chapter 2 introduces the Lewis and Vasishth (2005) model of cue-based retrieval parsing in sentence processing and its foundations in the cognitive architecture ACT-R (Anderson et al., 2004). This influential model and the principles of ACT-R are the basis for this thesis and the modeling presented in the following chapters.

In Chapter 3, the classical assumptions of cue-based retrieval in sentence processing, based mainly on ACT-R and the Lewis and Vasishth model, are assessed computationally with respect to a comprehensive literature review on retrieval interference in subject-verb and reflexive dependencies. It is shown that the cue-based retrieval account in its current form cannot explain several reported interference effects, such as (i) speed-ups observed in presence of a syntactically unlicensed distractor when the correct dependent is a full match to the retrieval cues and (ii) slowdowns when the correct dependent only partially matches the retrieval cues. It is demonstrated that these effects can be explained by two theoretical and independently motivated constructs: *Distractor prominence* accounts for the influence of experimental design on distractor activation, and *cue confusion* is a mechanism that underspecifies associations between certain features and retrieval cues according to the linguistic environment and possibly individual differences. The Lewis and Vasishth cue-based retrieval model is therefore extended to incorporate distractor prominence and cue confusion, and quantitative predictions are derived from this extended model. The extended model is shown to provide a better explanation of published results than the classical account. The contents of this chapter have been submitted as an article to the Journal of Memory and Language (Engelmann, Jäger, & Vasishth, 2015).

Chapter 4 establishes a basic model of the interaction between oculomotor control and parsing by implementing a linking hypothesis which is inspired by Reichle et al. (2009) and Mitchell et al. (2008) and states that short regressions mostly serve the function of compensating for slow post-lexical processing. First, the ACT-R eye movement control module EMMA is tested by replicating a corpus simulation on sentence reading of Salvucci (2001) within a newer ACT-R version. Then, the linking mechanism, which is termed Interface I: *Time Out* (after Mitchell et al's "Time Out hypothesis"), is implemented as a connection between EMMA and the Lewis and Vasishth (2005) sentence parser. In corpus-based simulations, the predicted influence of retrieval parsing difficulty and surprisal (Hale, 2001; Levy, 2008) on fixation measures is compared to empirical data.

Through the Time Out interface, EMMA's predictions are not only extended with post-lexically generated inter-word regressions, but also EMMA's performance is improved in all fixation measures. Furthermore the model predicts significant influences of surprisal and retrieval in both early and late measures similar to what has been found by Boston et al. (2011). The majority of the contents of this chapter have been published in Engelmann, Vasishth, Engbert, and Kliegl (2013).

Chapter 5 connects to the previous chapter in that it pursues to extend the basic eye-parser interaction model toward a near-complete set of interfaces and compares their predictions to examples from the literature. Interface II: *Reanalysis* triggers a regression whenever the parser needs to reanalyze previously built structure. A model combining Time Out and Reanalysis, without any parameter fitting, accurately predicts the results of Staub (2010b) on English relative clause processing, showing qualitatively different effects of memory retrieval and disconfirmed expectation in the same sentence.

Eye-parser Interface III: *Underspecification* ensures uninterrupted reading by abandoning non-vital, time-consuming parsing decisions. The time available for an attachment is, in the respective cases, constrained by eye movement control and available memory resources. A simulation of ambiguous adjunct attachments shows that different strategies of underspecification and reanalysis as a function of working memory capacity (von der Malsburg & Vasishth, 2013) emerge from a common underlying mechanism.

An account for the spill-over of post-lexical effects is offered by Interface IV: *Subvocalization.* It implements the possibility of storing a small number of words in the articulatory loop (Baddeley, 2003) for a short time while the parser is occupied with the integration of earlier material, thereby delaying integration processes and time-out regressions. The idea is demonstrated in a simple simulation reproducing spill-over effects in the study of Staub (2010a).

Chapter 6 finally summarizes the presented work and its implications. As an outlook for future work, it is furthermore speculated about the function and target-selection of long-range regressions and about a possible implementation of syntactic surprisal in ACT-R.

Chapter 2

Cue-Based Retrieval in Sentence Processing: The Lewis & Vasishth (2005) Model

Any comprehensive theory of sentence comprehension needs to explain the mechanisms behind the formation of dependencies between non-adjacent words such as a verb and its subject. This process necessarily involves storing and accessing information in working memory. There is a large body of evidence for a content-addressable memory architecture underlying human cognition in general (Anderson et al., 2004; Anderson & Lebiere, 1998; McElree, 2006; Ratcliff, 1978; Watkins & Watkins, 1975) and sentence processing in particular (Lewis & Vasishth, 2005; McElree, 2000; McElree, Foraker, & Dyer, 2003; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2011). In a content-addressable memory, a cue-based retrieval mechanism can activate certain items in parallel on the basis of how well their properties, i.e., their *features*, agree with a set of requirements, i.e. *cues*, which are determined by the type of dependency. This stands in contrast to search mechanisms which assume that items in memory are checked based on their location in memory, e.g., their serial order position (Berwick & Weinberg, 1984; Sternberg, 1966, 1969) or their position in a syntactic tree (Sturt, 2003). A cue-based model of sentence parsing has been described in Van Dyke and Lewis (2003), Lewis and Vasishth (2005), Lewis, Vasishth, and Van Dyke (2006), and Vasishth and Lewis (2006). Lewis and Vasishth (2005) implemented the model in the cognitive architecture *Adaptive Control of Thought Rational* (ACT-R, Anderson et al., 2004; Anderson & Lebiere, 1998) with the objective of grounding the means of language processing in general cognitive mechanisms. The parser uses rapid associative memory retrievals to form inter-word dependencies, incrementally building a structural sentence representation. The success and latency of the retrieval process depends on the activation of syntactic representations, which is affected by time-based decay and interference from similar items.

2.1 ACT-R

ACT-R is a comprehensive, implemented theory which integrates processes of working memory access, rule-guided behavior, learning, sensual input, and motor control. It is a constantly developing framework that incorporates findings from experimental work in various areas of cognitive psychology. ACT-R consists of a long-term memory of declarative and procedural knowledge and short-term buffers with limited capacity, representing a limited focus of attention (Cowan, 2001; McElree, 2006; Miller, 1956). Procedural knowledge is realized in the form of a production system Newell (1973, 1978) that consists of condition-action pairs that operate on short-term, or *working* memory. The contents of short-term memory buffers serve as conditions that trigger productions to fire. The result of a condition — the manipulated buffer contents — serve as condition for other productions. Hence, there is no supervised sequence of events; only condition-action specifications that operate autonomously by serially firing productions. At the same time, associated items are related by a mechanism of spreading activation that affects an individual item's activation dependent on the presence of related items. Memory items, so-called *chunks*, enter the buffers by being retrieved from declarative memory. Memory items are accessed by a cue-based retrieval mechanism on the basis of their activation, which is subject to decay, reactivation, similarity-based interference, and noise.

Next, the equations that underlie ACT-R content-addressable memory access will be explained in a — sometimes simplified — way as they are relevant for the model of sentence comprehension described in Lewis and Vasishth (2005).

In ACT-R, the probability and latency of retrieving a memory item is determined by its activation value. An item's activation fluctuates over time as the result of decay, reactivation, and noise. At the time of a retrieval request, a limited amount of activation spreads among all items in relation to their match with the retrieval specification, which is defined by a comparison of an item's *features* with the retrieval *cues*.

An item's final activation value is the sum of four components: A base-level B_i, which includes decay and frequency of use, the spreading activation S_i, which includes similarity-based interference, a penalty component P_i for mismatches with the retrieval specification, and a random noise component ε. The base-level activation B_i is computed from a *base-level constant* β_i and the item's history of use:

$$B_i = \ln(\sum_{j=1}^{n} t_j^{-d}) + \beta_i \tag{2.1}$$

where n is the number of times the item was accessed in memory, t_j is the time since the jth access, and d is the *decay parameter*. An item's activation decreases over time, with a decay parameter of 0.5 by default, and receives a reactivation boost when it is accessed.

At the time of a retrieval request, activation is spread from each retrieval cue to all matching items. This activation, however, is limited for each cue and distributed among the items that share the requested feature, i.e., the *competitors*. The number of items competing for activation from a certain cue is called the *fan*. An item with a high fan will thus receive less spreading activation than one with no competitors, i.e., it is inhibited by similarity-based interference or the so-called *fan effect*.[1] The spreading activation component S_i of item i is summed over all cues $j \in J$ in the retrieval specification:

$$S_i = \sum_j W_j S_{ji} \tag{2.2}$$

where W_j is the amount of activation from the cue j and S_{ji} is the strength of association between cue j and item i. S_{ji} is a function of the fan of item i for cue j:

$$S_{ji} = S - \ln(\mathit{fan}_{ji}) \tag{2.3}$$

where S is the *maximum associative strength* (MAS). The fan of item i for cue j is defined by the number of competing items in memory that match the feature associated with j: $\mathit{fan}_{ji} = 1 + \mathit{items}_j$.[2] As a consequence, the spreading activation component S_i increases the item's activation by Equation (2.2) for each matching retrieval cue and reduces its activation by Equation (2.3) for each distractor item that also (partially) matches the retrieval cues. Note that (2.3) is the core equation of similarity-based interference in ACT-R as it defines the influence of competitors on an item's activation at the time of retrieval.

The final activation component assigns a penalty for mismatches. Some activation is subtracted for each retrieval cue j that is not matched:

$$P_i = \sum_j PM_{ji} \tag{2.4}$$

P is the *mismatch penalty parameter* (MP). M_{ji} is the similarity between cue value j and the value in the corresponding slot of item i. The similarity M_{ji} is a value between 0 (identity) and -1 (maximum difference). This way, the more dissimilar a feature value of an item is to the cue value, the more activation is subtracted for this item. If, for example, one defines some similarity between the color values *red* and *orange*, orange items would be less penalized than, e.g., blue

[1] Note that the description of ACT-R for the present purpose is simplified to the way it was used in the model of Lewis and Vasishth (2005). In default ACT-R, spreading activation is not a property of retrieval cues per se. Rather, any buffer's content can spread activation to related items. Usually, the chunks in the goal buffer are the sources of spreading activation and, hence, of the fan effect. In the Lewis and Vasishth parser, the retrieval cues are always mirrored in the goal buffer, such that the model behaves as if spreading activation is specific to retrieval cues.

[2] In fact, all *slots* in all memory items containing the feature are counted. For simplicity, we assume here that a linguistically relevant item will have a certain feature in only one slot.

items when cueing for a *red*.

The item that is finally retrieved is the one with the highest activation. The time to retrieve an item i is a function of its activation A_i:

$$RT = Fe^{-(f \times A_i)} \tag{2.5}$$

where F is the *latency factor* (LF) and f the *latency exponent*. If no item has an activation above a certain threshold τ, retrieval fails. The duration of a failed retrieval is calculated by the same equation (2.5) with A_i substituted with τ.

2.2 Lewis & Vasishth (2005)

The computational model of parsing difficulty developed by Lewis and Vasishth (2005) adopts ACT-R's general principles, a limited focus of attention and cue-based retrieval of memory items subject to fluctuating activation as a function of decay and retrieval history and similarity-based retrieval interference. An essential property of cue-based retrieval in contrast to structural search is that there is no serial order information inherently available to the search mechanism (McElree, 2006; Ratcliff, 1978). Lewis and Vasishth (2005) argue that this serves the speed of language processing where most dependencies can be established without serial order information, purely based on the items features and recency in terms of activation decay over time. The lack of immediate serial order information in incremental sentence processing explains severe comprehension difficulty in cases where this information is needed like, e.g., in double center-embeddings such as Example (2):

(2) The book that the editor who the receptionist married admired ripped.

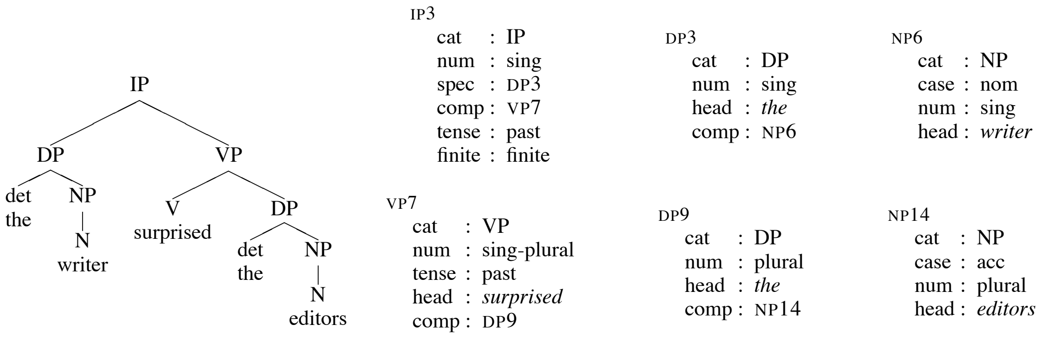

Figure 2.1: Figure 1 of Lewis and Vasishth (2005). Chunk representation.

The model of Lewis and Vasishth implements parsing knowledge in the form of production rules that incrementally build a structural representation in the fashion of a left-corner parser following X-bar rules (Chomsky, 1986). Figure 2.1 from Lewis and Vasishth (2005) shows how the a resulting sentence structure is represented in memory. Syntactic constituents are stored as single chunks being related to each other through feature slots for *specifier*, *complement*, and *head*. New structure is built at a new input word and then attached into previously built structure by retrieving a syntactic object that matches certain search cues like gender, number, syntactic category, and also information as to whether the relevant constituent is embedded or contains a gap waiting to be filled. The productions essentially operate on four buffers, each holding one chunk: A control buffer (the goal buffer), a lexical buffer holding the lexicon entry corresponding to the current word, a retrieval buffer holding the syntactic chunk retrieved from memory, and a buffer for creating new structure. The goal buffer contains some syntactic expectation in the form of a syntactic category that is necessary in order to complete the currently pursued structure. Figure 2.2 from Lewis and Vasishth (2005) illustrates the cycle carried out at each input word: First, the corresponding lexical entry is accessed in the lexicon in declarative memory. Based on the lexical entry and on the current goal category, the cues for retrieving a matching constituent are specified and retrieval is initiated. Finally, a new syntactic node is created and attached to the one retrieved. Attention is then sent to the next word. The essential step is working memory retrieval. Through ACT-R's independently motivated principles of cue-based working memory access, the simulations in Lewis and Vasishth (2005) provide quantitative predictions for effects of distance, structural interference, and embedding type in sentence comprehension.

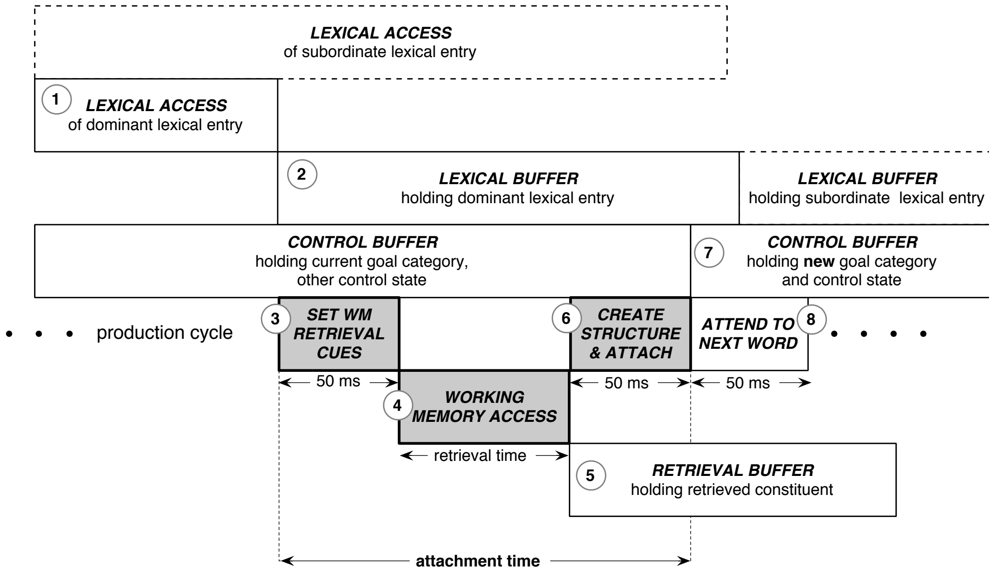

Figure 2.2: Figure 2 of Lewis and Vasishth (2005). Processing cycle.

The parsing architecture has been further used to model important aspects of sentence comprehension such as anti-locality (Vasishth & Lewis, 2006), intrusive interference in negative polarity constructions (Vasishth, Bruessow, Lewis, & Drenhaus, 2008), interference effects in reflexive processing (Jäger, Engelmann, & Vasishth, 2015; Parker & Phillips, 2014; Patil et al., 2012) and subject-verb processing (Dillon et al., 2013; Wagers et al., 2009), and impaired sentence comprehension in aphasia (Patil, Hanne, Burchert, De Bleser, & Vasishth, 2015).

Chapter 3

The Determinants of Interference: An Extended Model of Cue-Based Retrieval

The contents of this chapter have been submitted for publication in the *Journal of Memory and Language.*

3.1 Introduction

Ever since its conception, the cue-based retrieval model of Lewis and Vasishth (2005) and Lewis et al. (2006) has evolved as a standard reference for predictions related to cue-based retrieval in sentence processing. Many studies on interference effects in sentence processing refer to this model to evaluate their observations (e.g., Chen, Grove, & Hale, 2012; Cunnings & Sturt, 2014; Jäger, Benz, Roeser, Dillon, & Vasishth, 2015; Kush & Phillips, 2014; Van Dyke, 2007; Xiang, Dillon, & Phillips, 2009). Others have re-used it for simulating interference effects in subject-verb agreement (Dillon et al., 2013; Wagers et al., 2009) or in reflexive-antecedent dependencies (Dillon et al., 2013; Jäger, Engelmann, & Vasishth, 2015; Parker & Phillips, 2014; Patil et al., 2012). In this chapter, an implementation of the model is tested against commonly stated predictions as well as a comprehensive review of empirical studies. Because the ACT-R based Lewis and Vasishth (2005) model is widely used in sentence comprehension, the descriptions of cue-based retrieval in this chapter and the rest of the thesis are derived from this model (henceforth LV05), although it is just one out of different possible accounts of cue-based retrieval (e.g., the account of McElree et al., 2003 differs with respect to the relation between activation and retrieval speed).

A central prediction of cue-based retrieval is that items that are similar to the retrieval target in the sense that they share some features can interfere with the retrieval process even if they are structurally inaccessible: Similar items can slow down retrieval or occasionally be retrieved instead of the target. For example, in a reflexive-antecedent dependency such as (3)[1], Principle A of Binding Theory (Chomsky, 1981) only licenses *Bill* as an *accessible* antecedent of the reflexive *himself*, because it is the local c-commander. But a content-addressable memory retrieval mechanism would also consider the structurally *inaccessible* antecedent *Jane/John* because it is an animate noun phrase and thus matches some of the requirements for being an antecedent of *himself.*

(3) $[_{\text{Distr.}}$ Jane$^{-masc}_{-c\text{-}com}$/John$^{+masc}_{-c\text{-}com}]$ thought that $[_{\text{Target}}$ Bill$^{+masc}_{+c\text{-}com}]$ owed *himself*$\left\{\begin{smallmatrix} masc \\ c\text{-}com \end{smallmatrix}\right\}$ another opportunity to solve the problem.

In Example (3), the retrieval specification is shown as a set of cues in curly brackets behind the critical element that triggers retrieval. Match or mismatch of an element with a certain cue is represented by the name of the feature prefixed with a − or a +, respectively. Because the masculine feature of *John* matches the masculine cue of *himself*, *John* and *Bill* are harder to discriminate than *Jane* and *Bill.* This increased difficulty to discriminate is called *similarity-based interference* and, in LV05, causes increased retrieval latencies at the reflexive with the distractor *John* compared to the condition with *Jane.* A number of studies have found elevated reading times in the interference condition of sentences like (3), confirming the prediction of similarity-based interference (Badecker & Straub, 2002; Chen et al., 2012; Felser, Sato, & Bertenshaw, 2009; Jäger, Engelmann, & Vasishth, 2015).

Analogously, similarity-based interference is also expected in subject-verb dependencies. In (4)[2], *neighbor* is structurally unlicensed to control agreement with *was complaining* but is nevertheless assumed to interact with the retrieval process of the dependency completion mechanism at the verb. Confirmation for this has been found by Van Dyke and Lewis (2003), Van Dyke and McElree (2006), Van Dyke (2007), and Van Dyke and McElree (2011).

(4) The worker was surprised that the $[_{\text{Target}}$ resident$^{+anim}_{+locSubj}]$ who was living near the dangerous $[_{\text{Distr.}}$ warehouse$^{-anim}_{-locSubj}$/neighbor$^{+anim}_{-locSubj}]$ *was complaining*$\left\{\begin{smallmatrix} anim \\ locSubj \end{smallmatrix}\right\}$ about the investigation.

Throughout this chapter, we refer to the syntactically correct element of a dependency as *target* and to a syntactically unlicensed retrieval candidate as *distractor.* The observation of elevated reading times caused by the mechanism of similarity-based interference is called *inhibitory interference.* In terms of the ACT-R-based LV05 model, the inhibitory effect is explained by a competition between the target and the distractor for a limited amount of activation: Since the amount of activation associated with a retrieval cue is shared between all matching items, the presence of competitors in memory will reduces each item's activation. Since retrieval speed is a

[1] Badecker and Straub (2002)

[2] Van Dyke (2007)

function of an item's activation, reduced activation due to a cue-matching distractor results in a longer retrieval latency as compared to a condition without a cue-matching distractor.[3]

In addition to the prediction of elevated reading times, interference increases the probability of erroneously retrieving the partially matching distractor. These occasional *misretrievals* are predicted to cause incorrectly formed dependencies, affecting comprehension in the respective trials. In special situations, misretrievals of the distractor can lead to an observed speed-up in mean reading times. This is sometimes called *intrusion* and refers to cases where a distractor causes an ungrammatical sentence to be perceived as grammatical. A well-known example is the case of number attraction: In sentences like (5)[4], the plural distractor (*cabinets*) is predicted to *facilitate* processing, resulting in faster reading at the verb *were*. Several studies on number agreement have found this effect (Dillon et al., 2013; Lago, Shalom, Sigman, Lau, & Phillips, 2015; Pearlmutter et al., 1999; Tucker, Idrissi, & Almeida, 2015; Wagers et al., 2009).

(5) *The [$_{\text{Target}}$ key$^{-plur}_{+locSubj}$ to the [$_{\text{Distr.}}$ cabinet$^{-plur}_{-locSubj}$/cabinets$^{+plur}_{-locSubj}$] *were*$\{^{plur}_{locSubj}\}$ rusty from many years of disuse.

This facilitation is predicted for reflexive-antecedent dependencies, too, and has been observed, e.g., by King, Andrews, and Wagers (2012) in sentences similar to (6). Note that (6) is not actually ungrammatical, but only violates the *stereotypical* gender of *mechanic*, which is perceived as masculine. We therefore refer to sentences like (5) and (6) uniformly as *target-mismatch* conditions in contrast to sentences like (3) and (4), which we refer to as *target-match* conditions.

(6) The [$_{\text{Target}}$ *mechanic*$^{-fem}_{+c\text{-}com}$] who spoke to [$_{\text{Distr.}}$ John$^{-fem}_{-c\text{-}com}$/Mary$^{+fem}_{-c\text{-}com}$] sent a package to *herself*$\{^{fem}_{c\text{-}com}\}$ ...

The important characteristic of target-mismatch sentences such as (5) and (6) is that both the target and the distractor only partially match the retrieval cues and do not overlap in the features manipulated. E.g., the target *mechanic*$^{-fem}_{+c\text{-}com}$ in (6) only matches the structural requirement while the distractor *Mary*$^{+fem}_{-c\text{-}com}$ only matches the gender requirement. When target and distractor do not overlap in the manipulated feature in the distractor-match condition, no similarity-based interference is predicted. However, because both partially match the retrieval cues, the probability of erroneously retrieving the distractor is predicted to increase. On average, this causes shorter retrieval latencies in the distractor-match condition. This is because, with two nearly equally probable candidates, retrieval can be described similar to a race process (see, e.g., Van Gompel et al., 2001), retrieving whatever item is currently the highest activated and thus the fastest to retrieve (considering fluctuating activation due to noise). We refer to this

[3] Notice that findings of inhibitory interference have also been interpreted as reflecting *encoding interference*. Within a content-addressable framework, not only the retrieval process but also the encoding and maintenance of items in memory can be affected by the items' mutual similarity. E.g. in the content-addressable memory model proposed by Oberauer and Kliegl (2006), the activation level of a memory item decreases as a function of the number of features it shares with other items and as a function of the number of other items it shares features with. In the current work, we focus on the retrieval account and do not discuss encoding interference as this is theoretically orthogonal to cue-based retrieval interference.

[4] Pearlmutter, Garnsey, and Bock (1999)

speed-up as *facilitatory interference*.

Table 3.1: Mechanisms and predictions of the Lewis and Vasishth (2005) cue-based retrieval model (LV05) compared with observations from the literature review.

Target	**Classical LV05**		**Empirical findings**
	Mechanism	**Predictions**	
Match	SBI & misretrievals	(A) Inhibition	(A1) No effect (A2) Inhibition (A3) Facilitation
Mismatch	Misretrievals	(B) Facilitation	(B1) No effect (B2) Facilitation (B3) Inhibition

Table 3.2: Mechanisms and predictions of the extended cue-based retrieval model.

Target	**Extended model**		**Explanation**
	Mechanism	**Predictions**	
Match	SBI & misretrievals **increase with prominence**	(A1) No effect (A2) Inhibition (A3) Facilitation	Low *distractor prominence* Increased *distractor prominence* Very high *distractor prominence*
Mismatch	Misretrievals **& SBI by cue confusion**	(B1) No effect (B2) Facilitation (B3) Inhibition	Very low *distractor prominence* Increased *distractor prominence* High *cue confusion*

In sum, we define two mechanisms that are responsible for interference effects according to the classical LV05 cue-based retrieval theory: One mechanism is similarity-based interference (SBI), which refers to a reduction in activation in target and distractor due to an overlap in the manipulated feature. SBI causes inhibitory effects and is present only in target-match conditions. The second mechanism is the occasional misretrieval of the distractor. This decreases mean retrieval time and thus causes facilitatory effects. Misretrievals happen in both target-match and target-mismatch conditions, while in target-match conditions, the effect of SBI is usually assumed to be stronger.

Table 3.1 summarizes the mechanisms and predictions of LV05. The cue-based retrieval theory of sentence processing (in particular, with regard to the architecture of Lewis & Vasishth, 2005) is associated with the following predictions regarding retrieval latencies at the retrieval site: (A) *inhibitory interference* is predicted in *target-match* conditions due to similarity-based interference, and (B) *facilitatory interference* is predicted in *target-mismatch* conditions due to intrusion. By saying that cue-based retrieval is *associated* with these predictions, we mean that (A) and (B) are verbal simplifications of a model which exhibits much more variable behavior when deriving predictions from a computational implementation. The literature on interference in dependency processing shows a lot of variability which the statements (A) and (B) do not cover (see last column in Table 3.1). Apart from a large variability in the effect sizes, even the signs of the effects often contradict the predictions (A) and (B). For example, some studies report *facilitatory interference* in *target-match* conditions (Cunnings & Felser, 2013; Sturt, 2003), whereas other studies report *inhibitory interference* in *target-mismatch* conditions (Cunnings & Felser, 2013; Jäger, Engelmann, & Vasishth, 2015; Kush & Phillips, 2014), neither of which is expected under the assumptions of cue-based retrieval. Moreover, the observed patterns of effects differ between types of dependencies. For example, based on the frequent failure to find interference effects in reflexive processing, some researchers (Dillon et al., 2013; Kush & Phillips, 2014; Nicol & Swinney, 1989; Phillips, Wagers, & Lau, 2011; Sturt, 2003; Xiang et al., 2009) have suggested that the set of retrieval cues in this dependency type might be restricted to syntactic cues only or that syntactic retrieval cues at least have priority over semantic cues.

Some of the variability in the results might, however, be accounted for by systematic differences between studies. For example, the experimental design can affect how prominently the distractor is encoded in memory, which in turn can influence the amount of interference it causes. One possible factor of this type is the linear order of target and distractor. In retroactive interference designs, where the distractor has been read more recently than the target, the distractor would be more active in memory than in a proactive interference design. This is predicted under the assumption of a cue-based memory architecture with a decay component (Lewis & Vasishth, 2005). Using English reflexives, Cunnings and Felser (2013) tested this prediction by manipulating the linear order of target and distractor between two experiments with otherwise similar designs. In the retroactive design, they found an interference effect for participants with low working memory capacity in target-match conditions which had not been found in the proactive design.

In addition, there is evidence that grammatical role affects the availability of a noun phrase in the sense that, e.g., subjects are more salient or accessible than objects (Brennan, 1995; Chafe, 1976; Grosz, Weinstein, & Joshi, 1995; Keenan & Comrie, 1977). Thus, the grammatical role of the distractor could likewise affect the strength of interference such that a distractor in subject position, as in Example (8), evokes stronger interference than a distractor in object position as in (7).

(7) The surgeon who treated **Jonathan** had pricked himself with a used syringe needle.

(8) The tough soldier that **Fred** treated in the military hospital introduced himself to all the nurses.

This assumption has been considered, for example, in the experiments of Cunnings and Felser (2013) and Patil et al. (2012).

Finally, distractor prominence might also be influenced by discourse saliency, for instance if the distractor is *topic* (Ariel, 1990; Chafe, 1976; Du Bois, 1987, 2003; Givón, 1983; Grosz et al., 1995; Gundel, Hedberg, & Zacharski, 1993). Sturt (2003) and Cunnings and Felser (2013) used designs similar to (9), where the distractor noun phrase was introduced in a context sentence, making the distractor the discourse topic.

(9) **Jonathan** was pretty worried at the City Hospital. **He** remembered that the surgeon had pricked himself with a used syringe needle. There should be an investigation soon.

In the present work, we examine to what extent systematic differences between experiments such as distractor prominence, language, and the type of dependency can account for the variability of results reported for retrieval interference. We report the findings of a comprehensive literature review on interference in reflexive-antecedent dependencies, subject-verb number agreement, and other subject-verb dependencies. We argue that some variability in effect sizes, as well as the failure to find effects in various cases, and contradictory directions of effects could be systematically explained by differences in experimental preconditions between studies. We propose that two determinants are largely responsible for the variability in the literature: (1) *Distractor prominence*, which refers to the linear order of the antecedents, the structural position of the distractor, its discourse saliency, and the combinations thereof, affects the magnitude of interference. We show that, theoretically, a very high distractor prominence can even flip the usually inhibitory target-match interference effect to being facilitatory. (2) An independently motivated mechanism called *cue confusion* causes competition for activation even between conceptually different features, resulting in similarity-based interference in *target-mismatch* conditions.

We present a computational model that extends the cue-based retrieval model described in Lewis and Vasishth (2005) with a *prominence correction* and the principle of *associative cues*, thus predicting effects of distractor prominence and cue confusion on the strength and direction of interference. As summarized in Table 3.2, the extended model predicts that the magnitude of both inhibitory and facilitatory interference effects increases with distractor prominence (observations A1 vs. A2, B1 vs. B2) and target-match effects can be facilitatory for very high prominence values (A3). Furthermore, we propose that certain linguistic environments can induce cue confusion, which predicts inhibitory interference in target-mismatch conditions for these cases. The extended model is evaluated in quantitative simulations of the data from the literature review and compared to the standard LV05 model.

3.2 Literature Review

We reviewed 69 experiments examining interference in either reflexive-antecedent or subject-verb dependencies. The studies investigated are summarized in Tables 3.3 and 3.4. All studies have in common that the experimental manipulation targets the match of a certain retrieval cue (e.g., gender) with the target (i.e., the reflexive's structurally licit antecedent or the verb's subject, respectively) and with a distractor noun phrase. We collected the direction and magnitude of the observed effects as well as specifics of the experiments: Language, method, interference type (proactive vs. retroactive), and distractor position.

3.2.1 Method

Reported interference effects in Tables 3.3 and 3.4 represent the difference between the means of the distractor-match condition and the distractor-mismatch condition within target-match and target-mismatch conditions. A positive effect indicates inhibitory interference (slow-down) in the distractor-match condition whereas a negative effect reflects facilitatory interference (speed-up). An exception is the d' measure of SAT experiments, where a negative effect indicates inhibition. Whenever effect sizes were not directly reported by the authors, we derived them by subtracting the mean of the distractor-mismatch condition from the mean of the distractor-match condition. In case no condition means were provided in a publication, we only report the direction of the effect (inhibition versus facilitation) by labeling the effect as *inhib* or *facil* for a positive or negative interference effect, respectively. When separate condition means aggregated over participants and items were reported, we averaged these two values for the tables. Some studies manipulated another factor in addition to the target-match/mismatch and distractor-match/mismatch manipulations or manipulated the match of not only one but two retrieval cues. For the review, we disentangled these manipulations such that we report interference effects within each level of these additional factor separately or, in the case of multiple cues being manipulated, we report the effect caused by one cue and the effect caused by the other cue separately.

For each effect, we provide the region on which it was observed. The definition of the regions of interest varies across experiments. We have therefore decided to only distinguish between the critical and the post-critical regions. Any region containing the verb or the reflexive was labeled as the critical region, no matter how much material surrounding the verb or reflexive is included. The *two* regions directly following the critical region were uniformly labeled as post-critical, irrespective of their length. We collapsed the two post-critical regions because, across studies, the regions of interest differ in the number of words contained. Any effects further downstream in the sentence are not considered here. In the sentence classification experiments reported by Nicol, Forster, and Veres (1997) only total reading time of the whole sentence was measured. In this case we labeled the region of interest as *sent*.

We have further categorized the experiments by language, experimental method (including the dependent variable), retrieval cue under investigation, interference type (retroactive interference when the distractor followed the target and proactive when the distractor preceded the target) and the position of the distractor (e.g., subject, object, or memory load in a dual task paradigm). We restricted the review on effects reported for linguistically unimpaired, native, adult participants, not reporting studies on patients, children or second-language processing.

Note that the comparisons we apply to number agreement differ from the comparisons usually applied in this line of research: Most authors investigating number agreement were interested in agreement attraction effects, i.e., an effect of ungrammaticality being attenuated by the presence of a distractor that matches the verb's retrieval cues in the ungrammatical condition. Therefore, studies on number attraction mostly tested for main effects of grammaticality and number-match between the target and the distractor (rather than match with the retrieval cues) and, most importantly, for an interaction between the two. Since we are testing the theory of cue-based retrieval, for all dependency types we consistently report the effect of distractor match vs. mismatch with the verb and, respectively, with the reflexive within target-match and target-mismatch conditions separately. The main effect of grammaticality is not of primary interest. We therefore recoded the experimental conditions and the statistical comparisons such that they match the viewing angle of cue-based retrieval. This recoding of the comparisons will be explained in detail in the discussion of number agreement.

3.2.2 General Overview

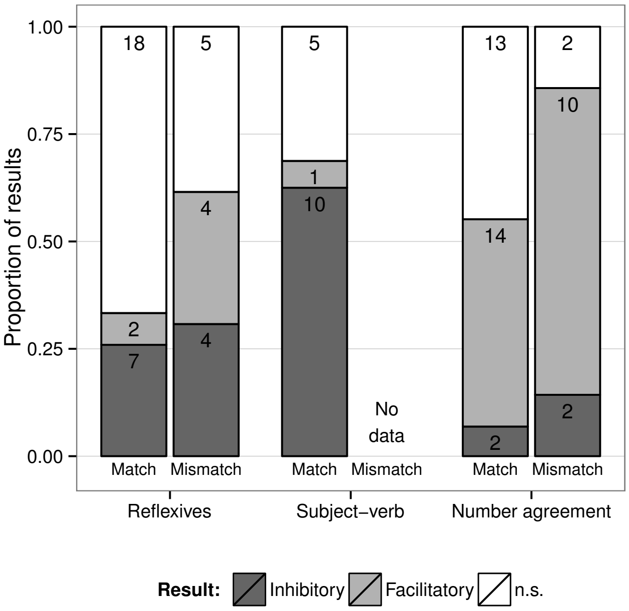

Figure 3.1: Proportion of studies reporting inhibitory and facilitatory effects (including marginal) by target-match and target-mismatch conditions. Proportion of facilitatory effects in light gray, proportion of inhibitory effects in dark gray. The absolute number of studies within each type of result is printed inside the bars.

Figure 3.1 plots the proportions of studies that reported inhibition, facilitation, or non-significance by dependency type and condition. With regard to subject-verb dependencies, we grouped the studies into two classes: Experiments studying number agreement and those that manipulated cues other than number. The reason for the separation of these studies should be obvious from

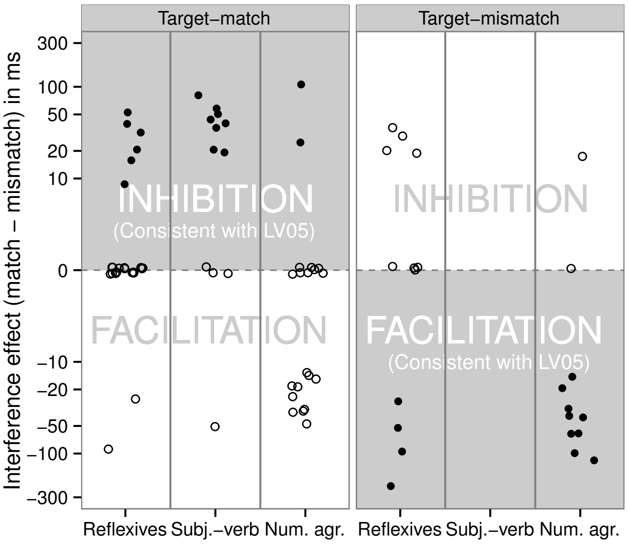

Figure 3.2: Overview of interference effects (distractor-match − distractor-mismatch) reported in the literature (only eye tracking and self-paced reading). The effects are represented in ms on a logarithmic scale. Effects above zero indicate inhibitory interference, effects below zero indicate facilitatory interference. Left panel: Target-match interference. Right panel: Target-mismatch interference. Black filled data points in areas with gray background are *consistent* with the commonly assumed predictions of the Lewis and Vasishth model (LV05). Non-filled circles represent data points that are apparently *inconsistent* with LV05.

looking at the figure: In target-match conditions, the results of the two groups seem to systematically contradict each other. Number agreement consistently exhibits facilitatory interference while the other group almost entirely shows inhibition. In general, effects in target-mismatch conditions seem to be easier to detect than in target-match conditions, an asymmetry that has been discussed by Wagers et al. (2009) in the context of number attraction. In target-match conditions of studies on reflexives, no effect was found in 18 out of 27 studies. Studies on non-number subject-verb dependencies detect effects more often, as there are 11 significant results in 16 experiments.

In order to provide an overview of effect sizes and a direct comparison with the predictions of cue-based retrieval, Figure 3.2 summarizes reported reading times of the eye tracking and self-paced reading studies included in the review. Included is one measure per study. The effect on the critical word is shown whenever there was one; the post-critical region is used otherwise. For eye-tracking studies, if possible, a measure that showed an effect in both target-match and target-mismatch conditions is plotted. The eye-tracking measure plotted is, if available, first-pass reading time (FPRT), otherwise total fixation time (TFT) or first-fixation duration (FFD). Studies on reflexive-antecedent dependencies show the most variability in the reported effects.

In target-match conditions, there are a number of studies that failed to find an effect. Six studies found inhibitory effects as predicted, and two reported facilitatory effects, contrary to the prediction of cue-based retrieval. In target-mismatch conditions, the picture is similar: Four studies failed to find an effect, four reported facilitatory interference, and four other studies found inhibition. An important observation is that, as mentioned above, in number agreement, nearly all effects are facilitatory, independent of whether the target matches the retrieval cues (grammatical) or not (ungrammatical).

The consistent facilitatory interference in target-match conditions of number agreement that was revealed by the recoding of comparisons is the opposite of prediction (A) of cue-based retrieval. We conclude that other mechanisms than cue-based retrieval must be responsible for the interference effects in number agreement. In contrast, the majority of the results of studies which investigated subject-verb dependencies without manipulating number as well as results of research on reflexives and reciprocals does not contradict the predictions of cue-based retrieval. Another interesting pattern we found in the review is that the variability in effect sizes, absent effects, and results that contradict the predictions (A) and (B) of cue-based retrieval seems to be related to study-specific variables such as interference type, distractor position, the language studied, and the type of dependency. The following sections present the review in detail.

Table 3.3: Reported interference effects in reflexive-antecedent comprehension, grouped for feature-match/mismatch of the target antecedent.

	Publication	Lang.	Method	Cue	Interf.	Distractor	Target-Match		Target-Mismatch	
					Type	Position	*Effect*	*AOI*	*Effect*	*AOI*
1	Xiang et al. '09	EN	ERP	gend	retro	subj.	- - -		*n.s.*	
2	Badecker&Straub '02 Exp5	EN	SPR	gend	pro	Gen.	*n.s.*		- - -	
3	Badecker&Straub '02 Exp6	EN	SPR	gend	pro	prep.obj.	*n.s.*		- - -	
4	Clifton et al. '99 Exp1	EN	SPR	num	retro	prep.obj.	*n.s.*		- - -	
5	Clifton et al. '99 Exp2	EN	SPR	gend,num	retro	prep.obj.	*n.s.*		- - -	
6	Clifton et al. '99 Exp3	EN	SPR	gend	pro	subj	*n.s.*		- - -	
7	Felser et al. 09 natives [*gend*]	EN	ET	gend	pro	subj.(topic)	*n.s.*		- - -	
8	Nicol&Swinney '89	EN	Primg	gend	pro	subj.,obj.	*n.s.*		- - -	
9	Dillon et al. '13	EN	ET	num	retro	obj.	*n.s.*		*n.s.*	
10	King et al. '12 [*adjacent*]	EN	ET	gend	retro	obj.	*n.s.*		*n.s.*	
11	Sturt '03 Exp2	EN	ET	gend	retro	obj.(topic)	*n.s.*		*n.s.*	
12	Cunnings&Felser '13 Exp1	EN	ET (FPRT)	gend	pro	subj.(topic)	*n.s.*		(−25)	post
13	Parker&Phillips '14 Exp1	EN	ET (TFT)	num/gend	pro	subj.	*n.s.*		≈ −250	crit
14	Exp2		(TFT)	num/anim			*n.s.*		≈ −170	crit
15	Exp3		(TFT)	gend/anim			*n.s.*		≈ −250	crit
16	Cunnings&Sturt '14 Exp1	EN	ET (FPRT)	gend	pro	subj.(topic)	*n.s.*		(+**39**)	post
17	Jäger et al. (2015) Exp1	CN	ET (FPRT)	anim	retro	subj.	*n.s.*		+**19**	crit
			(RBRT)				*n.s.*		+**18**	crit
			(FFD)				*n.s.*		+**13**	crit
			(RPD)				*n.s.*		+**10**	crit
18	Cunnings&Felser '13 Exp2 [*lWM*]	EN	ET (FPRT)	gend	retro	subj.(topic)	−**49**	crit	*n.s.*	
			(FFD)				−**24**	crit	(+**22**)	crit
19	Sturt '03 Exp1	EN	ET (RRT)	gend	pro	subj.(topic)	−**97**	post	*n.s.*	
20	Badecker&Straub '02 Exp3	EN	SPR	gend	pro	subj.	+42	post	- - -	
21	Felser et al. 09 natives [*c-com*]	EN	ET (RPD)	c-com	pro	subj.(topic)	+29	crit	- - -	
22	Jäger et al. (2015) Exp2	CN	ET (FPRT)	anim	pro	3x memory	+15	crit	- - -	
			(RBRT)				+19	crit	- - -	
			(RPD)				+67	crit	- - -	
			(TFT)				+42	crit	- - -	
23	Patil et al. (2012)	EN	ET (FPRP)	gend	retro	subj.	+6.74%	crit	*n.s.*	
			(regr.-cont. FFD)				(≈ +20)	crit	(≈ −50)	crit
	Prepositional Reflexives									
24	Clackson et al. '11 Exp2 [*adults*]	EN	VW	gend	pro	subj.(topic)	*n.s.*		- - -	
25	King et al. '12 [*non-adjacent*]	EN	ET (FPRT)	gend	retro	prep.obj.	*n.s.*		≈ −95	crit
26	Clackson&Heyer '14	EN	VW (target ident.)	gend	pro	subj.(topic)	*inhib*	200-600 ms	- - -	
	Possessive Reflexives									
27	Chen et al. '12 [*local*]	CN	SPR	anim	retro	subj.	*n.s.*		- - -	

Continued on next page

Table 3.3 – continued from previous page

	Publication	**Lang.**	**Method**	**Cue**	**Interf.**	**Distractor**	**Target-Match**		**Target-Mismatch**	
					Type	**Position**	*Effect*	*AOI*	*Effect*	*AOI*
28	Chen et al. '12 [*non-local*]	CN	SPR	anim	retro	subj.	+9	post	- - -	
	Reciprocals									
29	Badecker&Straub '02 Exp4	EN	SPR	num	pro	subj.	+48	post	- - -	
30	Kush&Phillips '14	HI	SPR	num	retro	prep.obj.	*n.s.*		(+**30**)	post

Note. Studies are sorted by direction of effect (target-mismatch within target-match). Numbers represent effect size in milliseconds if no other unit is shown. Positive values indicate inhibition, negative values indicate facilitation. Marginal effects are in parentheses. Non-significant results are denoted with *n.s.*; '- - -' means that the respective manipulation was not tested. Effects not consistent with assumptions about cue-based retrieval based on Lewis and Vasishth (2005) are shown in bold. The experiments are classified by language (EN = English, CN = Mandarin Chinese, HI = Hindi), experimental method (SPR = self-paced reading, ET = eye tracking while reading, VW = visual world eye tracking, ERP = event-related potentials, Primg: cross-modal priming), retrieval cues that are examined (gend = gender, num = number, anim = animacy), dependency type (reflexive, prepositional reflexive, possessive reflexive, reciprocal), interference type (proactive vs. retroactive), and by syntactic position of the distractor (subject, object, genitive attribute, sentence external memory load, discourse topic). For reading studies, the interest area (AOI) labeled n refers to the word position of the reflexive/reciprocal. For eye-tracking experiments, Measure indicates the dependent variable in which the effect was observed.

Table 3.4: Reported interference effects for subject-verb dependency comprehension.

	Publication	Lang.	Method	Interf.	Distractor	Singular Verb				Plural Verb			
				Type	Position	gram		ungram		gram		ungram	
						Effect	*AOI*	*Effect*	*AOI*	*Effect*	*AOI*	*Effect*	*AOI*
	NUMBER AGREEMENT												
31	Franck et al. ’15 [*Compl*]	FR	SPR	pro	obj	*n.s.*		- - -		- - -		- - -	
32	Dillon et al. ’13 Exp1	EN	ET	retro	obj	*n.s.*		- - -		- - -		−118	crit
33	Kaan ’02	NL	ERP (500-700 ms)	retro	obj	*n.s.*		**+6.41 µV**	crit	*n.s.*		*n.s.*	
34	Lago et al. ’15 Exp1	SP	SPR	pro	subj	*n.s.*		- - -		- - -		−39	post
35	Lago et al. ’15 Exp2	EN	SPR	pro	subj	*n.s.*		- - -		- - -		−36	post
36	Lago et al. ’15 Exp3B	SP	SPR	pro	subj	*n.s.*		- - -		- - -		−21	post
37	Nicol et al. ’97 Exp5 [*lw-attchmt*]	EN	sent. class.	retro	obj[lwRC]	*n.s.*		- - -		- - -		- - -	
38	Pearlmutter ’00 Exp2 [*1st distr.*]	EN	SPR	retro	PP,PP	- - -		- - -		**−19**	crit	- - -	
39	Wagers et al. ’09 Exp2	EN	SPR	pro	subj	*n.s.*	(post)	- - -		- - -		−58	post
40	Wagers et al. ’09 Exp3	EN	SPR	pro	subj	*n.s.*		*n.s.*		*n.s.*		(*facil*)	post
41	Acuña et al. ’14	SP	ET (TFT)	retro	PP	**−15**	crit	- - -		**−32**	crit	- - -	
			(Reg in)			**−5%**	crit	- - -		**−1%**	crit	- - -	
			(cumRT)			**−25**	crit	- - -		**−59**	crit	- - -	
			(FPRP)			**−5%**	crit	- - -		**−5%**	crit	- - -	
			(FFD)			**−7**	post	- - -		*n.s.*		- - -	
42	Lago et al. ’15 Exp3A	SP	SPR	pro	subj	**−12**	post	- - -		- - -		−15	post
43	Nicol et al. ’97 Exp1	EN	maze	retro	PP	**−70**	crit	- - -		*n.s.*		- - -	
44	Nicol et al. ’97 Exp2	EN	sent. class.	retro	PP	**−124**	sent	- - -		*n.s.*		- - -	
45	Nicol et al. ’97 Exp4	EN	sent. class.	retro	PP	**−60**	sent	- - -		- - -		- - -	
46	Nicol et al. ’97 Exp5 [*hg-attchmt*]	EN	sent. class.	retro	obj[hgRC]	**−67**	sent	- - -		- - -		- - -	
47	Pearlmutter et al. ’99 Exp1	EN	SPR	retro	PP	**−35**	crit	- - -		- - -		**+19**	crit
						−36	post	- - -		- - -		−26	post
48	Pearlmutter et al. ’99 Exp2	EN	ET (TFT)	retro	PP	**−49**	post	- - -		- - -		−106	crit
												−44	post
			(FPRT)			**−36**	post	- - -		- - -		*n.s.*	
			(FPRP)			*n.s.*		- - -		- - -		−15%	crit
49	Pearlmutter et al. ’99 Exp3	EN	SPR	retro	PP	**−36**	crit	- - -		+24	post	- - -	
50	Pearlmutter ’00 Exp1 [*2nd distr.*]	EN	SPR	retro	PP,PP	**−23**	crit	- - -		- - -		- - -	
51	Tucker et al. ’15	AR	SPR	retro	obj	**−14**	post	- - -		- - -		−57	crit
52	Wagers et al. ’09 Exp4	EN	SPR	retro	PP	**−17**	crit	- - -		- - -		−32	post
53	Franck et al. ’15 [*RC*]	FR	SPR	pro	obj	≈ +100	crit	- - -		- - -		- - -	

Continued on next page

Table 3.4 – continued from previous page

	Publication	Lang.	Method	Interf.	Distractor	Singular Verb				Plural Verb			
				Type	Position	gram		ungram		gram		ungram	
						Effect	*AOI*	*Effect*	*AOI*	*Effect*	*AOI*	*Effect*	*AOI*
	STRUCTURAL CUES (SUBJ.)												
54	Van Dyke '07 Exp1 [*LoSem*]	EN	SPR	retro	PP/subj	*n.s.*		- - -		- - -		- - -	
55	Van Dyke '07 Exp2 [*LoSem*]	EN	ET (FPRT)	retro	PP/subj	+37	crit	- - -		- - -		- - -	
			(RPD)			+140	crit	- - -		- - -		- - -	
						+395	post	- - -		- - -		- - -	
			(TFT)			+180	crit	- - -		- - -		- - -	
			(FPRP)			+6%	crit	- - -		- - -		- - -	
56	Van Dyke '07 Exp3 [*LoSem*]	EN	ET (FPRT)	retro	PP/subj	+20	crit	- - -		- - -		- - -	
			(RPD)			+40	crit	- - -		- - -		- - -	
57	Van Dyke & Lewis '03 Exp4	EN	SPR	retro	PP/subj	+56	crit	- - -		- - -		- - -	
	SEMANTIC CUES												
58	Van Dyke & McElree '11 Exp2a [*pro*]	EN	SAT (d')	pro	obj	*n.s.*		- - -		- - -		- - -	
59	Van Dyke & McElree '11 Exp2a [*retro*]	EN	SAT (d')	retro	obj	*n.s.*		- - -		- - -		- - -	
60	Van Dyke & McElree '11 Exp2b [*pro*]	EN	ET (TFT)	pro	obj	*n.s.*		- - -		- - -		- - -	
61	Van Dyke & McElree '11 Exp2b [*retro*]	EN	ET (TFT)	retro	obj	*n.s.*		- - -		- - -		- - -	
62	Van Dyke '07 Exp3 [*LoSyn*]	EN	ET (RPD)	retro	PP	(−**54**)	post	- - -		- - -		- - -	
63	Van Dyke '07 Exp1 [*LoSyn*]	EN	SPR	retro	PP	+54	crit	- - -		- - -		- - -	
64	Van Dyke '07 Exp2 [*LoSyn*]	EN	ET (FPRT)	retro	PP	+44	post	- - -		- - -		- - -	
			(RPD)			(+23)	crit	- - -		- - -		- - -	
						+235	post	- - -		- - -		- - -	
65	Van Dyke & McElree '06	EN	SPR	pro	3x memory	+38	crit	- - -		- - -		- - -	
66	Van Dyke & McElree '11 Exp1a [*pro*]	EN	SAT (d')	pro	subj	*inhib* −0.16	crit	- - -		- - -		- - -	
67	Van Dyke & McElree '11 Exp1a [*retro*]	EN	SAT (d')	retro	subj	*inhib* −0.27	crit	- - -		- - -		- - -	
68	Van Dyke & McElree '11 Exp1b [*pro*]	EN	ET (TFT)	pro	subj	+20	crit	- - -		- - -		- - -	
69	Van Dyke & McElree '11 Exp1b [*retro*]	EN	ET (TFT)	retro	subj	+81	crit	- - -		- - -		- - -	

Note. Studies are sorted by direction of effect in grammatical singular verb conditions and, within that, alphabetically. Numbers represent effect size in milliseconds if no other unit is shown. Positive values indicate inhibition, negative values indicate facilitation. Marginal effects are in parentheses. Non-significant results are denoted with *n.s.*; '- - -' means that the respective manipulation was not tested. Effects not consistent with classical assumptions about cue-based retrieval based on Lewis and Vasishth (2005) are shown in bold. The experiments are classified by language (AR = Arabic, EN = English, FR = French, NL = Dutch, SP = Spanish), experimental method (SPR = self-paced reading, ET = eye tracking while reading, ERP = event-related potentials, SAT = speed-accuracy trade-off, sentence classification, maze task), retrieval cues that are examined, interference type (proactive vs. retroactive), and by syntactic position of the distractor (subject, object, prepositional phrase, sentence external memory load).

3.2.3 Reflexive-Antecedent Dependencies

Table 3.3 summarizes 30 experiments on interference in the processing of reflexives. Most studies investigating interference effects in reflexive-antecedent dependencies employed a feature-match/mismatch design as the one shown in (10), borrowed from Sturt (2003). In this design, a cue-relevant feature (e.g., gender, number, or animacy) is manipulated on the structurally licit antecedent (a noun phrase c-commanding the reflexive inside the reflexive's binding domain), which we refer to as *target*, and on a noun phrase which is in a structurally inaccessible position, i.e., not c-commanding the reflexive, which we refer to as *distractor*.

(10) a. *Target-match; distractor-match*
The surgeon$^{+masc}_{+c\text{-}com}$ who treated Jonathan$^{+masc}_{-c\text{-}com}$ had pricked himself$\{^{masc}_{c\text{-}com}\}$

b. *Target-match; distractor-mismatch*
The surgeon$^{+masc}_{+c\text{-}com}$ who treated Jennifer$^{-masc}_{-c\text{-}com}$ had pricked himself$\{^{masc}_{c\text{-}com}\}$

c. *Target-mismatch; distractor-match*
The surgeon$^{-fem}_{+c\text{-}com}$ who treated Jennifer$^{+fem}_{-c\text{-}com}$ had pricked herself$\{^{fem}_{c\text{-}com}\}$

d. *Target-mismatch; distractor-mismatch*
The surgeon$^{-fem}_{+c\text{-}com}$ who treated Jonathan$^{-fem}_{-c\text{-}com}$ had pricked herself$\{^{fem}_{c\text{-}com}\}$

For the review, we considered reflexives in direct object position, possessive reflexives such as the Chinese *ziji-de* ("himself/herself"-[*possessive*]), reflexives inside a prepositional phrase (e.g., *of himself*) and reciprocals (*each other*). All of these anaphors require a c-commanding antecedent inside their binding domain, i.e., they follow Principle A of Binding Theory (Chomsky, 1981). We do not consider reflexives inside a so-called picture-noun phrase because they presumably differ in their syntactic properties from other types of reflexives. In contrast to number agreement, in reflexives, the target-mismatch conditions are not always ungrammatical. Some studies on English reflexives use nouns which have a stereotypical gender. Hence, the gender manipulation on the target leads to a violation of the *stereotypical* gender in mismatch conditions, but not to ungrammaticality. It has been argued that the violation of the stereotypical gender results in a perceived ungrammaticality at least in the initial stage of processing the reflexive (Sturt, 2003).

3.2.3.1 Results for Reflexive-Antecedent Dependencies

The literature on interference in reflexive processing shows a lot of variability. Moreover, except for Cunnings and Felser (2013) and Patil et al. (2012), we did not find any studies showing effects in both target-match and target-mismatch conditions within the same experiment. Most studies found an effect in only one of these configurations. However, a large number of studies failed to find an effect in target-match conditions in eye tracking while reading (Sturt, 2003, Experiment 2; Felser et al., 2009; Cunnings & Felser, 2013, Experiment 1; Cunnings & Sturt, 2014; Dillon et al., 2013; Jäger, Engelmann, & Vasishth, 2015; King et al., 2012; Parker & Phillips, 2014, Experiment 1), self-paced reading (Badecker & Straub, 2002, Experiment 5 and 6; Clifton, Frazier, & Deevy, 1999; Chen et al., 2012; Kush & Phillips, 2014), cross-modal priming (Nicol & Swinney, 1989), and visual world eye tracking (Clackson, Felser, & Clahsen, 2011, Experiment 2).

Absent or small effects in target-match conditions are predicted by the so-called "grammatical asymmetry" (Wagers et al., 2009): Based on experimental results on number agreement, Wagers et al. proposed an asymmetric relation between interference effects in target-match sentences and in target-mismatch sentences, which they motivated with assumptions about cue-based retrieval. They argued that, in sentences where the target item fully matches the retrieval cues, its activation would strongly out-compete that of partially matching competitors, such that a distractor has only a minor influence compared with target-mismatch conditions. Consequently, interference effects in target-match sentences would be absent or much smaller than in target-mismatch sentences where the activations of target and distractor are more equal. To remain consistent with our terminology, we will call this the *match-mismatch asymmetry*. We believe that a match-mismatch asymmetry is not a robust prediction in reflexive-antecedent dependencies, since there are a number of studies that did find interference effects in target-match conditions.

As predicted by cue-based retrieval, effects found in target-match conditions are very consistently reported as inhibitory (Badecker & Straub, 2002, Experiment 3 and 4; Felser et al., 2009, natives, syntactic manipulation; Jäger, Engelmann, & Vasishth, 2015, Experiment 2; Chen et al., 2012; Clackson & Heyer, 2014; Patil et al., 2012). Two publications, however, report a significant facilitation in target-match conditions: In their Experiment 2, Cunnings and Felser (2013) found facilitatory interference in first-pass reading times and first-fixation durations only for readers with low working memory span; Sturt (2003) found a facilitation in post-critical rereading times in their Experiment 1. Notably, in all nine studies reporting a target-match interference effect, the distractor was in subject position or there were multiple distractors presented as memory load in a dual task paradigm. This could indicate that a more prominent distractor or having multiple distractors increases the chances of finding an effect.

Target-mismatch conditions have been tested much more rarely than target-match conditions. Out of 13 studies, four experiments failed to find an effect: Dillon et al. (2013); King et al. (2012), and Sturt, 2003's Experiments 1 and 2. Three eye-tracking studies found the facilitation that is expected under the assumptions of cue-based retrieval (Cunnings & Felser, 2013, marginal in Experiment 1; Parker & Phillips, 2014; Patil et al., 2012, marginal; King et al., 2012). Inconsistent with the predictions of cue-based retrieval, some experiments report *inhibitory* interference in target-mismatch conditions: Using eye tracking, an inhibitory effect was found by Jäger, Engelmann, and Vasishth (2015), Experiment 1, for Mandarin, Cunnings and Sturt (2014) (marginal post-critical effect in English materials), and Cunnings and Felser (2013), Experiment 2 (marginal effect for low-span readers of English); Kush and Phillips (2014) found a marginal inhibitory effect at the post-critical region in self-paced reading of Hindi reciprocals.

3.2.3.2 Discussion of Reflexive-Antecedent Dependencies

The results of the literature review on reflexive-antecedent dependencies are mixed and therefore it is hard to draw a clear conclusion in favor or against cue-based retrieval. The observed facilitation in target-match conditions and the inhibition in target-mismatch conditions are incompatible with the standard cue-based retrieval architecture in the sense of Lewis and Vasishth (2005), since these effects have the opposite sign of what this model predicts. Furthermore, neither a purely structural search account nor a cue-based account that prioritizes structural cues is able to explain the presence of interference effects observed in several experiments. The fact that there are many non-significant results in target-match conditions might be reconcilable under cue-based retrieval if a match-mismatch asymmetry is assumed as Wagers et al. (2009) have proposed for agreement interference. Indeed, the majority of experiments that showed an effect in target-mismatch conditions did not show one in target-match conditions. The data is, however, not conclusive, since most experiments that showed target-match effects did not test target-mismatch conditions.

The results for reflexive-antecedent dependencies indicate that interference effects are affected by the grammatical role of the distractor (subject vs. object) and also by discourse topicality. However, the results do not show strong evidence for a relation between effect sizes and linear order of the antecedents (pro- vs. retroactive). An exception is a controlled comparison by Cunnings and Felser (2013), who found support for an influence of linear order on interference effects in participants with low working memory.

A unique approach of manipulating distractor prominence by decreasing the target's match has been followed by Parker and Phillips (2014). In several eye-tracking experiments, they compared reflexive-antecedent interference in the usual target-mismatch conditions such as (11c and d) with conditions where the target had *two* mismatching features as in (11e and f). They found that the two-feature mismatch conditions showed an interference effect while none was observed in the one-feature mismatch conditions.

(11) a. *Target-match; distractor-match*
The librarian$^{+fem}_{+sing}$ said that the **schoolgirl**$^{+fem}_{+sing}$ reminded herself$\{^{fem}_{sing}\}$ about the book.

b. *Target-match; distractor-mismatch*
The janitor$^{-fem}_{+sing}$ said that the **schoolgirl**$^{+fem}_{+sing}$ reminded herself$\{^{fem}_{sing}\}$ about the book.

c. *Target-1-feature-mismatch; distractor-match*
The librarian$^{+fem}_{+sing}$ said that the **schoolboy**$^{-fem}_{+sing}$ reminded herself$\{^{fem}_{sing}\}$ about the book.

d. *Target-1-feature-mismatch; distractor-mismatch*
The janitor$^{-fem}_{+sing}$ said that the **schoolboy**$^{-fem}_{+sing}$ reminded herself$\{^{fem}_{sing}\}$ about the book.

e. *Target-2-feature-mismatch; distractor-match*
The librarian$^{+fem}_{+sing}$ said that the **schoolboys**$^{-fem}_{-sing}$ reminded herself$\{^{fem}_{sing}\}$ about the book.

f. *Target-2-feature-mismatch; distractor-mismatch*
The janitor$^{-fem}_{+sing}$ said that the **schoolboys**$^{-fem}_{-sing}$ reminded herself$\{^{fem}_{sing}\}$ about the book.

Parker and Phillips explained their results with a priority of syntactic cues over semantic cues. The authors showed with a computational model that the observed pattern was predicted when assuming a weighting of structural cues which is about three times the weighting of semantic cues. We will return to this experiment in the simulations section and the General Discussion.

3.2.4 Subject-Verb Dependencies

Studies investigating retrieval interference in subject-verb dependencies have a similar design as the experiments examining reflexive-antecedent dependencies: The subject (i.e., the retrieval target) and a structurally inaccessible distractor noun either match or mismatch a certain non-structural retrieval cue. However, in the literature on the processing of subject-verb dependencies, one can distinguish two lines of research which developed largely independently from each other: Studies on number attraction (e.g., Nicol et al., 1997; Pearlmutter et al., 1999) and studies on retrieval interference in subject-verb dependencies in general (e.g., Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2011).

The idea of number attraction in subject-verb dependencies originates in sentence production experiments like Bock and Miller (1991) that observed agreement attraction errors as shown in Example 12. Participants made agreement errors., i.e., erroneously produced a verb that mismatched the subject's number feature, more often when the first NP (the subject) was singular and the second NP was plural. This effect was explained by "upward percolation" of the distractor's plural feature, thus overwriting the singular number of the head subject (Bock & Eberhard, 1993; see also Eberhard, 1997; Franck, Vigliocco, & Nicol, 2002; Vigliocco, Butterworth, & Semenza, 1995).

(12) The key^{+sing} to the cabinets^{+plur} *was*/ **were* rusty from many years of disuse.

Nicol et al. (1997) followed by Pearlmutter et al. (1999) investigated whether agreement errors are unique to production or whether a similar phenomenon could be observed in sentence comprehension. They tested whether the presence of a number-matching distractor reduces the ungrammaticality effect induced by a number mismatch between the verb and its subject. Pearlmutter et al. (1999) used self-paced reading and eye tracking to test sentences as shown in (13) and tested for main effects of grammaticality (13a and b vs. 13c and d) and gender overlap between the subject and the distractor (13a and d vs. 13b and c) and their interaction (13a and c vs. 13b and d). They replicated the effect that had been found in production: Participants read the singular verb more slowly when the distractor was plural as in (13b). The effect was reversed, however, in ungrammatical conditions (c) and (d) where a plural verb was used with a singular head noun. Here, the plural distractor had a facilitatory effect. Both effects are compatible with the feature-overwriting account of Bock and Eberhard (1993): In both cases, the dominant plural feature overwrites the number of the head noun, leading to a perceived mismatch (13b vs. a) or match (13c vs. d) of the head noun with the verb.

(13) a. *Target-match (grammatical), singular verb; distractor-match / number-match*
The key$^{+sing}_{+locSubj}$ to the cabinet$^{+sing}_{-locSubj}$ was$\{{}^{sing}_{locSubj}\}$ rusty from many years of disuse.

b. *Target-match (grammatical), singular verb; distractor-mismatch / number-mismatch*
The key$^{+sing}_{+locSubj}$ to the cabinets$^{-sing}_{-locSubj}$ was$\{{}^{sing}_{locSubj}\}$ rusty from many years of disuse.

c. *Target-mismatch (ungrammatical), plural verb; distractor-match / number-mismatch*
The key$^{-plur}_{+locSubj}$ to the cabinets$^{+plur}_{-locSubj}$ were$\{{}^{plur}_{locSubj}\}$ rusty from many years of disuse.

d. *Target-mismatch (ungrammatical), plural verb; distractor-mismatch / number-match*
The key$^{-plur}_{+locSubj}$ to the cabinet$^{-plur}_{-locSubj}$ were$\{{}^{plur}_{locSubj}\}$ rusty from many years of disuse.

In the line of research targeting subject-verb dependencies other than number agreement, originating in Van Dyke and Lewis (2003)'s work, the author's were interested in interference effects as evidence for cue-based retrieval. In contrast to the first line of research, none of these studies investigated the number feature but other (potential) retrieval cues. For example, Van Dyke (2007) manipulated *animacy* (14b vs. a) and grammatical features like *subject/case* (14c vs. a).

(14) The worker was surprised that the resident$^{+anim}_{+subj}$

a. *LoSyn/LoSem*
who was living near the dangerous warehouse$^{-anim}_{-subj}$

b. *LoSyn/HiSem*
who was living near the dangerous neighbor$^{+anim}_{-subj}$

c. *HiSyn/LoSem*
who said that the warehouse$^{-anim}_{+subj}$ was dangerous

d. *HiSyn/HiSem*
who said that the neighbor$^{+anim}_{+subj}$ was dangerous

was complaining$\{{}^{anim}_{subj}\}$ about the investigation.

Although these two research lines addressed different research questions, the experimental manipulations they used are largely identical: Both use a feature match-mismatch design similar to the design used in the experiments on reflexive processing. However, the studies differ in their condition labeling and in the coding of the statistical comparisons. We therefore have relabeled the experimental conditions and report statistical comparisons in a way that the results are comparable across studies and also between subject-verb and reflexive-antecedent dependencies. The details of the recoding are explained below.

3.2.4.1 Recoding of Comparisons in Number Agreement

The comparisons used for number agreement are not consistent across studies. Most studies, however, report effects of grammaticality, number-match between target and distractor, and their interaction. The term *number-match* thereby refers to the match between target and distractor and not distractor and verb. Furthermore, the applied tests often involve the direct comparison of verbs in different number. For all studies we discuss in this literature review, we consistently compared a distractor match with a mismatch with respect to the verb's retrieval cues within target-match (grammatical) and target-mismatch (ungrammatical) conditions. We also did separate comparisons for singular verbs and for plural verbs.

Table 3.5 provides a complete example (taken from Wagers et al., 2009) of grammatical and ungrammatical conditions with both singular and plural verbs, labeled according to NP number match and verb-distractor match. Our comparisons with respect to this example are shown in Table 3.6. In grammatical conditions with singular verbs, we subtracted the distractor-mismatch condition [NP$_{\text{sg}}$-NP$_{\text{pl}}$-V$_{\text{sg}}$] from the distractor-match condition [NP$_{\text{sg}}$-NP$_{\text{sg}}$-V$_{\text{sg}}$] (a − b). Similarly, in grammatical conditions with plural verbs, we computed (c) − (d). In ungrammatical conditions, we computed (e) − (f) and (g) − (h). An interaction between grammaticality and gender overlap often reported in agreement attraction, e.g., (a) and (g) vs. (b) and (h), corresponds to a main effect of interference (i.e., an effect of distractor-match) in our coding.

The recoding of comparisons provides comparability not only between the agreement literature and the reflexive literature, but also *within* the literature on subject-verb dependencies, since the recoded comparisons correspond to those used in the second line of research on subject-verb dependencies (e.g., Van Dyke, 2007; Van Dyke & McElree, 2006, 2011).

Table 3.5: Full set of conditions for number agreement (Wagers et al., 2009).

	NP num	**Verb–distr.**
Target-match, Singular verb		
(a) The musician$^{+sing}_{-locSubj}$ who the reviewer$^{+sing}_{+locSubj}$ praises$\{^{sing}_{locSubj}\}$ …	match	**match**
(b) The musicians$^{-sing}_{+locSubj}$ who the reviewer$^{+sing}_{+locSubj}$ praises$\{^{sing}_{locSubj}\}$ …	mismatch	mismatch
Target-match, Plural verb		
(c) The musicians$^{+plur}_{-locSubj}$ who the reviewers$^{+plur}_{+locSubj}$ praise$\{^{plur}_{locSubj}\}$ …	match	**match**
(d) The musician$^{-plur}_{-locSubj}$ who the reviewers$^{+plur}_{+locSubj}$ praise$\{^{plur}_{locSubj}\}$ …	mismatch	mismatch
Target-mismatch, Singular verb		
(e) The musician$^{+sing}_{-locSubj}$ who the reviewers$^{-sing}_{+locSubj}$ praises$\{^{sing}_{locSubj}\}$ …	mismatch	**match**
(f) The musicians$^{-sing}_{-locSubj}$ who the reviewers$^{-sing}_{+locSubj}$ praises$\{^{sing}_{locSubj}\}$ …	match	mismatch
Target-mismatch, Plural verb		
(g) The musicians$^{+plur}_{-locSubj}$ who the reviewer$^{-plur}_{+locSubj}$ praise$\{^{plur}_{locSubj}\}$ …	mismatch	**match**
(h) The musician$^{-plur}_{-locSubj}$ who the reviewer$^{-plur}_{+locSubj}$ praise$\{^{plur}_{locSubj}\}$ …	match	mismatch

Table 3.6: Recoded comparisons we used for the literature review of studies on number attraction with respect to the example in Table 3.5.

	Our comparisons (distr.-match – distr.-mismatch)
Target-match	
Singular verb	(a) – (b)
Plural verb	(c) – (d)
Target-mismatch	
Singular verb	(e) – (f)
Plural verb	(g) – (h)

3.2.4.2 Results for Number Agreement

In the review of the literature on interference in subject-verb dependencies in Table 3.4, we report effects for singular and plural verbs separately. This distinction was not applied in the summary of the literature on reflexives, because only very few studies examine plural reflexives. More importantly, in reflexives, there is no indication that the processing of plural reflexives such as *themselves* is qualitatively different from the processing of singular reflexives such as *himself.* In subject-verb agreement, in contrast, it has been claimed that the agreement process is inherently different when using plural distractors compared to singular distractors. This difference, the so-called "number asymmetry" (Wagers et al., 2009), has been attributed to the morphological markedness of English plurals compared to the unmarked singular forms. In English reflexives, by contrast, the singular-plural distinction is a lexical one. Most of the studies on number agreement focused on the processing of singular verbs in grammatical configurations (target-match) and of plural verbs in ungrammatical configurations (target-mismatch). Singular verbs in target-mismatch were only tested in an ERP study by Kaan (2002) and in self-paced reading by Wagers et al. (2009). In fact, these two are the only studies that tested all four configurations that are shown in Table 3.5.

In general, the effects in number agreement collected in Table 3.4 show a striking consistency across languages and experimental methodologies. In target-match conditions with singular verbs, almost exclusively facilitatory interference has been found across experimental methods (self-paced reading, eye tracking, maze task, and a sentence classification) in English (Nicol et al., 1997; Pearlmutter, 2000; Pearlmutter et al., 1999; Wagers et al., 2009), Spanish (Acuña–Fariña, Meseguer, & Carreiras, 2014; Lago et al., 2015), and Arabic (Tucker et al., 2015). In eye-tracking experiments, the facilitatory effect was reliable across reading measures (Acuña–Fariña et al., 2014; Pearlmutter et al., 1999). The locus of the effect was the critical verb or the region following it, or both. In Nicol et al. (1997), the effect was observed in total reading times of the whole sentence. As the only exception to the pattern of facilitatory effects, Franck, Colonna, and Rizzi (2015) report inhibitory interference in target-match conditions of singular verbs in a self-paced reading experiment in French with proactive interference from a distractor that is the head of a relative clause.

In most of the studies reporting interference in singular target-match conditions, the distractor noun was dominated by a prepositional phrase, retroactively interfering with the subject. Wagers et al. (2009) have noted that these cases, especially where distractor and verb are adjacent, potentially represent spillover effects from the number manipulation of the distractor (see General Discussion for details on this). However, three studies show that number-matching distractors in other syntactic positions also induce interference effects: Nicol et al. (1997) observed retroactive facilitatory interference from a relative clause object in their Experiment 5, Lago et al. (2015) report in Experiment 3A *proactive* facilitatory interference from a distractor which is in the subject position of the matrix clause, and Franck et al. (2015) found *inhibitory* interference proactively from a distractor that is the head of a relative clause.

In plural verbs, only seven experiments tested grammatical (target-match) configurations (Acuña–Fariña et al., 2014; Kaan, 2002; Nicol et al., 1997; Pearlmutter, 2000; Pearlmutter et al., 1999; Wagers et al., 2009). Facilitatory interference was observed in two experiments testing Spanish (Acuña–Fariña et al., 2014) and English (Pearlmutter, 2000) materials on the critical region whereas inhibitory interference was observed in a single experiment testing English at the post-critical region (Pearlmutter et al., 1999). All three experiments tested the number feature in a retroactive interference configuration with the distractor being contained in a PP. In Pearlmutter (2000), this PP contained two distractors, but the facilitatory effect was solely caused by the first distractor. In target-match conditions with a plural verb, the effect of plural complexity would be different: In contrast to singular verbs, the predicted number effect for plural verbs is inhibitory, because here the distractor is plural in the distractor-match condition. Hence, retrieval interference and plural complexity make opposite predictions. However, the available data for plural target-match conditions is sparse and inconclusive with respect to the dominating direction of effects.

In target-mismatch conditions of singular verbs, the data is extremely sparse: Only Kaan (2002) reported an increased positivity in event-related potentials in the 500-700ms time window testing Dutch materials. The only other study testing this configuration (Wagers et al., 2009) did not find significant effects in eye tracking.

In target-mismatch conditions of plural verbs, facilitatory interference was observed in English (Dillon et al., 2013; Pearlmutter et al., 1999; Wagers et al., 2009), Spanish (Lago et al., 2015), and Arabic (Tucker et al., 2015) in both retro- and proactive interference configurations. The structural position of the distractor does not seem to have as much influence in these cases as it did in the grammatical singular-verb conditions. Moreover, in contrast to the facilitation observed in grammatical conditions of singular verbs, the locus of the effect in ungrammatical plural conditions appeared to be somewhat delayed in general: In only three experiments was facilitation observed at the critical region containing the verb (Dillon et al., 2013; Pearlmutter et al., 1999; Tucker et al., 2015) whereas in eight experiments, the effect reached significance only at the post-critical region. In Experiment 2 of Pearlmutter et al. (1999), an inhibitory effect at the critical region turned into a significant facilitation at the post-critical region. The only study that tested target-mismatch conditions with plural verbs and did not find an effect is Kaan (2002). One study not mentioned so far tested Dutch materials in singular-verb target-match and plural-verb target-mismatch conditions using event-related potentials (Severens, Jansma, & Hartsuiker, 2008). However, the interference effect with respect to our comparison coding cannot be derived from the data provided in the paper, since distractor-match and distractor-mismatch conditions are analyzed at different time windows and electrodes.

3.2.4.3 Results for Non-Number Subject-Verb Dependencies

Experiments studying cues other than number exclusively examined the processing of singular verbs in target-match configurations, and were only conducted in English. The respective studies are summarized in Table 3.4. For this line of research, no recoding of the experimental comparisons was necessary. However, some of the experiments manipulated a second factor in addition to the interference manipulation. Van Dyke and McElree (2011) manipulated interference type (proactive versus retroactive) and Van Dyke (2007) manipulated the retrieval cue under examination (semantic cue vs. syntactic cue, see Example 14 above). In our review of Van Dyke and McElree (2011), we report effects within proactive and retroactive constructions, respectively. For Van Dyke (2007) we report the effect of the semantic manipulation within the *LoSyn* (low syntactic interference) conditions and the syntactic manipulation within the *LoSem* (low semantic interference) conditions.

In contrast to experiments on number agreement, studies manipulating other features consistently report inhibitory rather than facilitatory interference in target-match conditions with singular verbs (Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006, 2011). This effect was observed in both retro- and proactive interference configurations and was consistent across methodologies (self-paced reading, eye tracking and speed-accuracy trade-off) and, in the case of eye-tracking experiments, reached significance across various dependent variables. The locus of the effect was mostly the critical verb, in several cases the effect also spilled over to the post-critical region, and in Van Dyke (2007), Experiment 3 (eye tracking), it was already significant at the pre-critical region. The position of the distractor did not affect the direction of the effect: interference is reported from sentence external memory load words (Van Dyke & McElree, 2006), structurally inaccessible subject noun phrases (Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2011), and noun phrases dominated by a PP (Van Dyke, 2007). Note, however, that Van Dyke and McElree (2011) conducted four experiments (proactive and retroactive, eye tracking and speed-accuracy trade-off) with the distractor in object position and did not find any effect of interference. This could indicate that the grammatical role of the distractor has some influence on the strength of interference. Only in one experiment did a distractor matching the semantic requirements of the verb lead to a processing speed-up (*facil-*

itation); in a single eye-tracking measure (regression-path duration) on the post-critical region (Van Dyke, 2007, Exp. 3, semantic manipulation). However, this effect only reached significance in the by-participants analysis and was not present in the by-items analysis. Hence, we conclude that this isolated facilitatory effect is likely to be a Type I error.

3.2.4.4 Discussion of Subject-Verb Dependencies

The literature on number agreement reveals a large body of evidence for facilitatory interference caused by a distractor which matches the number cue on the verb in target-match (grammatical) configurations of singular verbs and in target-mismatch (ungrammatical) configurations of plural verbs. In target-match (grammatical) configurations of plural verbs, not much data exist, but the evidence presented so far also points to facilitatory interference. Plural target-match and singular target-mismatch conditions have rarely been tested for the reason that the original production studies on number attraction did not show effects when the distractor was singular, as is the in these configurations.

An interesting finding of our literature review on subject-verb dependencies is that manipulating the number feature has different consequences than manipulating other features: While interference in target-match conditions is *facilitatory* in number agreement, interference is *inhibitory* in other cue manipulations. Moreover, the match-mismatch asymmetry (proposed by Wagers et al., 2009 as "grammatical asymmetry") does not seem to be a robust pattern, since effects have been found in ungrammatical *and* grammatical conditions. However, our review indicates that effects in ungrammatical conditions might be indeed stronger and hence easier to detect than those in grammatical conditions: While all 12 experiments testing ungrammatical sentences found an effect at least in one of the ungrammatical configurations, in grammatical sentences almost half of the experiments failed to find significant effects (12 vs. 14). As for the "number asymmetry" (Wagers et al., 2009), our review is inconclusive because conditions with singular distractors (singular verb/ungrammatical and plural verb/grammatical) have only rarely been tested. For grammatical plural verb conditions, we found an equal number of studies reporting a null result and studies reporting a significant interference effect. In singular target-match conditions of both number agreement and other subject-verb dependencies there are indications that interference type (retro- vs. proactive) and structural position of the distractor influence the strength of effects.

The predictions of the LV05 cue-based retrieval architecture are consistent with the results of studies on subject-verb dependencies investigating features other than the number feature. It is also able to explain the facilitatory effects caused by distractor-match in plural *target-mismatch* (ungrammatical) conditions in number agreement. However, it cannot explain the facilitatory effects of number interference in *target-match* (grammatical) conditions in singular or plural verbs. Most importantly, the current ACT-R model is unable to explain the apparent asymmetry between studies on number agreement and studies on other subject-verb dependencies. In order to achieve a deeper understanding of this asymmetry between retrieval cues, further research is needed on i) retrieval cues other than number in target-mismatch conditions of singular verbs and in both target-match and target-mismatch conditions of plural verbs and ii) on languages other than English. Moreover, more research is needed in target-mismatch (ungrammatical) configurations of singular verbs.

3.2.5 Summary of the Literature Review

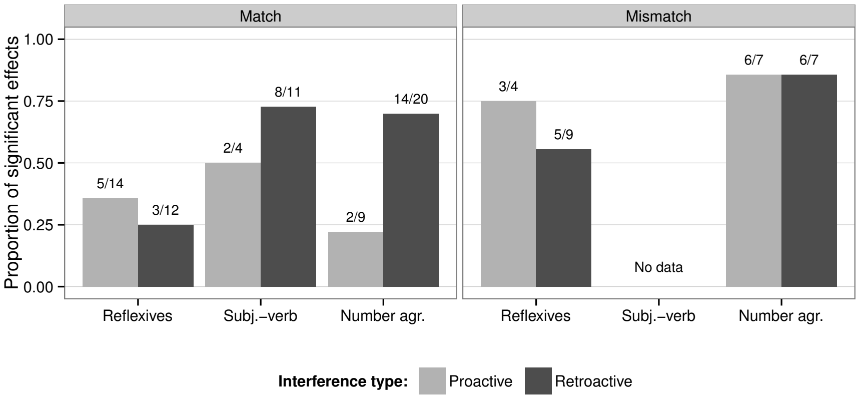

Figure 3.3: Proportion of reported effects (including marginal) by linear order of target and distractor (proactive vs. retroactive), showing that linear order has an influence mainly in target-match conditions of subject-verb dependencies and number agreement. Numbers above the bars indicate the number of significant studies and the total number of studies in the respective configuration.

The review reveals considerable variation in reported effects. Experiments on subject-verb agreement manipulating the number feature seem to stand out in that their effects are very homogeneous and, crucially, in that their target-match interference effects are consistently facilitatory. The consistent facilitation in target-match conditions is not only the opposite of what would be expected under cue-based retrieval but is also inconsistent with the majority of the target-match effects in all other studies. This asymmetry points toward a qualitative difference between the mechanisms underlying number agreement and other dependencies. In ungrammatical conditions, on the other hand, effects and explanations in terms of agreement attraction are compatible with cue-based retrieval: In ungrammatical conditions, cue-based retrieval predicts that a distractor that matches the verb is erroneously retrieved in some trials, resulting in an observed speed-up in mean reading times. This predicts the proposed illusion of grammaticality induced by, e.g., the presence of a plural distractor in the ungrammatical combination of a singular subject and a plural verb.

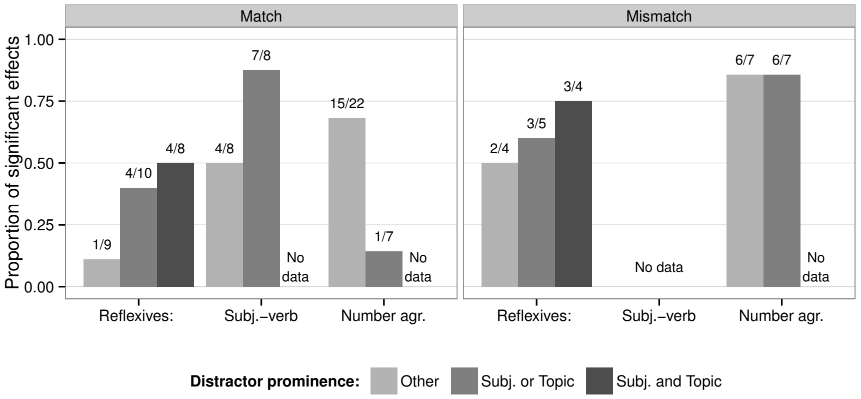

Figure 3.4: Proportion of reported effects (including marginal) by distractor's discourse prominence (subject and/or topic), showing that prominence increases the proportion of observed effects in reflexives and subject-verb dependencies but not in number agreement. Numbers above the bars indicate the number of studies reporting an effect and the total number of studies in the respective configuration.

Wagers et al. (2009) and others claim that number attraction can be explained by cue-based retrieval. However, the direction of the effects in target-match conditions speaks against such an explanation as no facilitation would be predicted in the presence of a matching distractor. A potential way to account for the results according to Wagers et al. (2009) is to say that accidental misretrievals of a *mismatching* distractor NP in the grammatical case lead to frequent reanalysis that increases reading times. When both NPs match in number in the distractor-match condition, a retrieval of the distractor is not as easily detectable as an error, resulting in fewer cases of reanalysis. With this additional assumption, the facilitatory interference in target-match conditions can be explained. However, this assumption would have to be made specifically for number agreement while it is unnecessary in most other types of dependencies. An alternative explanation is that the subject's number will be *misrepresented* or *overwritten* in the presence of a distractor with a different number (Bock & Eberhard, 1993; Nicol et al., 1997; Pearlmutter et al., 1999). This theory correctly predicts the results: In the target-match case, a mismatching distractor renders the sentence ungrammatical, making the distractor-match condition easier. In the target-mismatch (ungrammatical) conditions, a cue-matching distractor would also facilitate processing by changing the perceived number of the subject, making it grammatical. Suggested mechanisms that lead to faulty representations of the subject's number are feature percolation through a hierarchical tree representation (Franck et al., 2002; Nicol et al., 1997; Staub, 2009; Vigliocco et al., 1995) or spreading activation between the NPs' number features (Eberhard, Cutting, & Bock, 2005). Note, however, that the first account, feature percolation, can only explain effects in complex subjects such as a prepositional phrase where the distractor is hierarchically below the target. In relative clause constructions with proactive interference, this is not the case (Wagers et al., 2009). Furthermore, both theories assume that feature-overwriting occurs when the subject is singular and the distractor is plural, because plural is considered to be more strongly *marked* than singular (Lau, Rozanova, & Phillips, 2007; Lehtonen, Niska, Wande, Niemi, & Laine, 2006; New, Brysbaert, Segui, Ferrand, & Rastle, 2004). However, more data is needed in order to conclude whether this plural markedness is present not only in production but also in comprehension.

Note that other researchers have claimed previously that the processing of reflexives depends on a slightly different mechanism than the processing of agreement (Dillon et al., 2013; Kush & Phillips, 2014; Nicol & Swinney, 1989; Parker & Phillips, 2014; Phillips et al., 2011; Sturt, 2003; Xiang et al., 2009). Their reasoning rested on the observation that effects seem to be harder to detect in reflexives than in agreement. They proposed that in both cases cue-based retrieval is involved, but in reflexive processing structural cues have priority over semantic cues. We, too, propose a difference between both dependency types, however, in a different way: In order to explain target-match effects in agreement processing, an additional assumption (or mechanism different from cue-based retrieval) is needed, which is not necessary in any other dependency we looked at. We will come back to this in the General Discussion. An exhaustive discussion of the potential mechanisms behind number attraction is, however, beyond the scope of this thesis. For a clearer understanding of the memory processes that are involved in number attraction, further research is necessary that can disentangle encoding interference from retrieval interference. For the simulations we report below, we focused on the literature on reflexive-antecedent dependencies and on subject-verb dependencies not involving number manipulation.

Recall that the cue-based retrieval theory in sentence processing is associated with the following predictions: (A) inhibitory effects are predicted in target-match conditions due to similarity-based interference, and (B) facilitatory interference is expected in target-mismatch conditions due to intrusion. Inconsistent with (A) and (B), the literature review showed that *facilitatory* effects have occasionally been observed *target-match* conditions (A3) and *inhibition* has been observed in *target-mismatch* conditions (B3) (cf. Figure 3.1). Additionally, in both reflexives and subject-verb dependencies, there seems to be a trend toward a match-mismatch asymmetry proposed by Wagers et al. (2009) and others: There are many studies that have failed to find an effect in target-match conditions (A1).

Both the probability of observing an effect and the direction of target-match effects seem to be related to *distractor prominence* in terms of its structural position (subject vs. object), its discourse saliency (e.g., topic), and the linear order of the target and distractor (retro- vs. proactive).[5] There are, however, differences between dependency types. As Figure 3.3 shows, linear order of target and distractor is important only in subject-verb dependencies, especially in num-

[5] For our further discussion, we assume that there is no correlation between sample size and distractor prominence in the reviewed experiments. This assumption seems trivial but is crucial because, otherwise, an increased sample size would be an alternative explanation for an increased probability of finding a significant effect in experiments with a prominent distractor.

ber agreement. Distractor prominence, on the other hand, increases the effect in reflexives and in subject-verb dependencies that do not include number agreement, as illustrated in Figure 3.4. The fact that number agreement patterns differently from the other two dependencies on several levels supports the assumption that number agreement relies on different mechanisms.

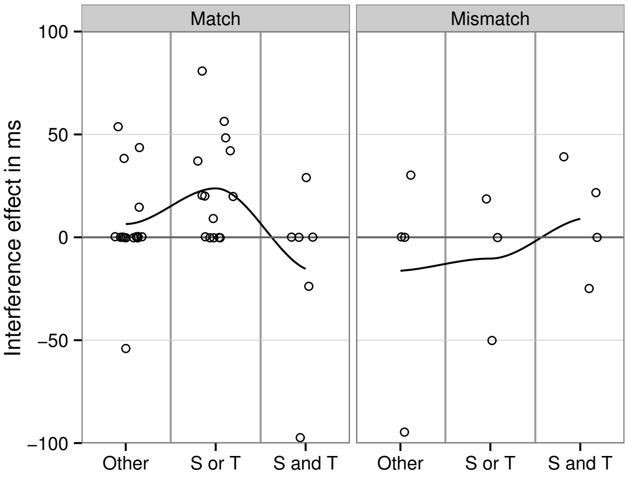

Figure 3.5: Interference effects by discourse prominence class for eye tracking and self-paced reading studies on reflexives and subject-verb dependencies, excluding studies on number agreement. Discourse prominence classes for the distractor are *Other*: neither subject nor topic; *S or T*: subject *or* topic; *S and T*: subject *and* topic. Lines represent locally fit polynomials (LOESS).

In Figure 3.5, reported effects for the same three prominence classes as in Figure 3.4 are plotted for reflexive-antecedent and subject-verb dependencies without number agreement. The means of the three prominence classes indicate that having the distractor in subject position or topicalized increases the inhibitory interference effect in target-match conditions. However, when the distractor is very prominent, being a topicalized subject, the mean of target-match interference effects is facilitatory. This is only a visual pattern, of course, since the data is too sparse and exhibits too much variance for strong statistical inferences. This is especially true for target-mismatch conditions, because these were not tested in subject-verb dependencies.

The four inhibitory effects reported in target-mismatch conditions (B3), which are important because they contradict classical assumptions about cue-based retrieval in sentence processing, are spread over all three prominence classes. Thus, there is no pattern showing a possible relation between these results and distractor prominence. Instead, we propose as an explanation the concept of *cue confusion*. This concept predicts that inhibitory interference can occur in target-mismatch conditions in special linguistic environments like reciprocals (Kush & Phillips, 2014) and Mandarin reflexives (Jäger, Engelmann, & Vasishth, 2015) or for participants with low working memory capacity (Cunnings & Felser, 2013).

In the following sections, we develop an extended model of cue-based retrieval which predicts effects of distractor prominence and cue confusion and compare the predictions with the results in the literature.

3.2.6 Predictions of ACT-R

It is essential to validate theoretical predictions in an implemented computational model. Verbal statements like (A) the prediction of inhibitory effects in target-match conditions and (B) the prediction of facilitatory effect in target-mismatch conditions are only abstractions of the potentially complex mathematical description of a model, and some interactions between model variables might not be obvious without simulations. We therefore evaluate the verbal predictions (A) and (B) with respect to the LV05 model computationally.

We test to what extent the observations (A1-3) and (B1-3) from the literature review could be accounted for in a principled way by the LV05 model. We therefore generated predictions over a range of parameter combinations. In order to investigate the relation between predicted effects and distractor prominence, we manipulated the distractor's base-level activation. For reasons of speed, simplicity, and transparency of our simulations, we implemented the retrieval process by means of the equations explained above in R (R Core Team, 2014). A simple R model was built to simulate a retrieval operation where two items, a target and a distractor, are available. The model simulates a standard four-condition design similar to Example 10. Target and distractor each hold two features which can match or mismatch with the respective cues of the retrieval specification. The first cue corresponds to structural accessibility and is always matched by the target and never matched by the distractor. In the target-mismatch condition, the target does not match the second cue. The distractor matches the second cue in the distractor-match condition. In the distractor-mismatch condition, the distractor does not match any of the retrieval cues. Predictions for interference effects were derived by running the model for 3000 iterations and subtracting the mean retrieval latency of the distractor-mismatch conditions from the mean latency of the distractor-match conditions within target-match and target-mismatch conditions.

3.2.6.1 General Predictions

Figure 3.6 plots interference effects (mean latency difference between distractor-match and distractor-mismatch conditions) predicted by ACT-R for a range of parameter values that should sufficiently cover the model's range of possible predictions. Values above zero indicate *inhibitory* interference (slow-downs), values below zero indicate a *facilitatory* effect (speed-up). With equal base-level activation of target and distractor, the simulation results roughly match the standard predictions: ACT-R predicts that (A) for target-match conditions only positive values should occur

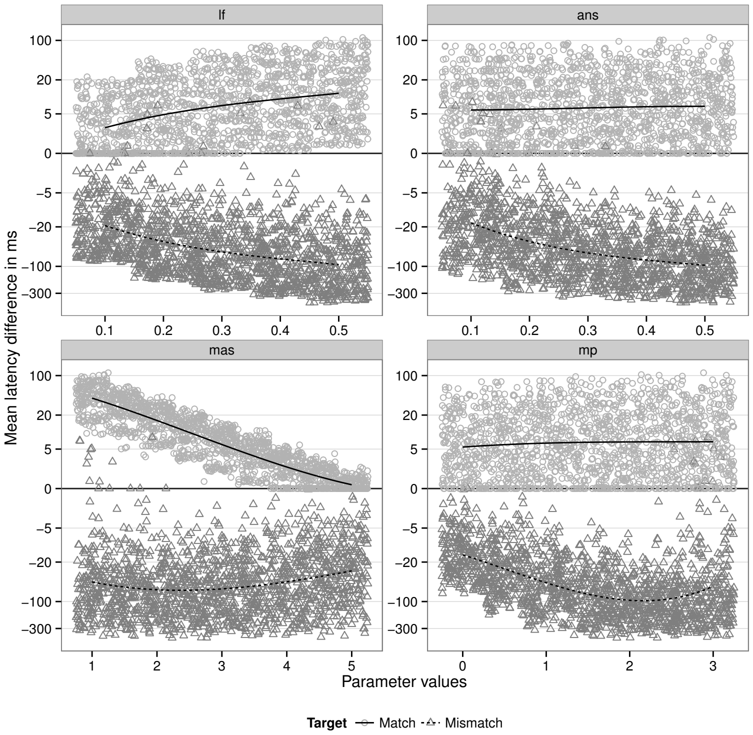

Figure 3.6: Interference effects (distractor-match − distractor-mismatch) predicted by ACT-R for 1575 parameter combinations in the following ranges: *latency factor* LF $= [0.1, 0.5]$, *noise* ANS $= [0.1, 0.5]$, *maximum associative strength* MAS $= [1, 5]$, *mismatch penalty* MP $= [0, 3]$. Retrieval threshold was at -1.5, distractor base-level activation at 0, other parameters at their ACT-R defaults. Y-axis is on a logarithmic scale.

due to similarity-based interference (the *fan effect*), and (B) for target-mismatch conditions only negative values should occur due to misretrievals of the distractor.

Prediction (A) strictly holds in Figure 3.6: For all parameter manipulations, target-match predictions (circles) are above zero. However, there are 13 positive values in target-mismatch conditions (triangles above zero). These values are related to a comparatively high *mismatch penalty* MP (mean $= 2.92$) and they account for about about 0.8 percent of the predictions in target-mismatch conditions. However, their mean size is negligible with 3.4 and their 95% confidence intervals (not shown in the plot) all include zero. We therefore conclude that these values do not represent a systematic prediction of inhibitory interference in target-mismatch conditions. Hence, both predictions (A) and (B) are confirmed by the simulations.

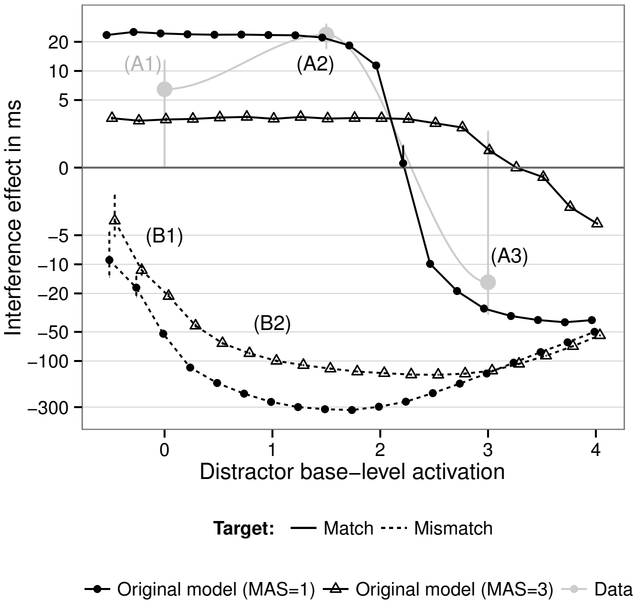

Figure 3.7: Relation between the distractor base-level activation B_{Distr} and the predicted interference effect (on a logarithmic scale) when target base-level is zero. Two models are compared: Original ACT-R with *maximum associative strength* parameter MAS = 1 (circles) and a model with MAS increased to 3 (triangles). Solid lines represent target-match conditions; dashed lines represent target-mismatch conditions. The light gray line represents the mean of empirical target-match effects by distractor prominence as a smoothed curve connecting the data means in the three prominence classes as shown in Figure 3.5. Error bars represent standard errors.

3.2.6.2 Distractor Activation

Figure 3.7 plots the predicted interference effects for target-match and target-mismatch conditions in relation to distractor activation. The original model with default parameters is shown as circles.

The literature review indicated that the probability of finding an effect is related to the grammatical position of the distractor, its discourse saliency, and the linear order of distractor and target. We look at the effect of distractor activation on *target-mismatch* conditions first (dashed lines). Effects in target-mismatch conditions are close to zero for a very low distractor base-level activation. This predicts that target-mismatch effects are hard to detect in constructions where the distractor is not prominent (B1). With increased distractor activation, the facilitatory target-mismatch effect increases due to a higher probability of misretrievals, accounting for observation (B2). The facilitatory effect reaches its maximum when the distractor's final activation at the time of retrieval (*base-level* + *spreading activation* + *mismatch penalty*) is equal to the target's final activation (in the figure at a base-level value of about 1.5). Above this point, the facilitatory effect decreases because, as the distractor activation deviates from the target activation, retrieval speed more and more depends on the distractor activation only and less on the dynamics between both target and distractor.

We now look at the effect of distractor activation on *target-match* conditions (solid lines) in Figure 3.7. The figure shows a stable prediction of inhibitory interference in target-match conditions at base-level values below 2 (A2). At a distractor base-level of 2, the inhibitory effect starts decreasing and finally turns into a facilitatory effect at a base-level of 2.5 and greater. The reason for this is the following: At very high base-level activations, the final distractor activation at retrieval time including spreading activation and mismatch penalty exceeds the target activation such that the distractor will be retrieved instead of the target in most trials of the distractor-match condition. In these trials, retrieval will be fast, since the distractor is highly activated, which leads to a facilitation in comparison with the distractor-mismatch condition. This prediction strictly follows from an activation-based retrieval mechanism but may not be very obvious until an implemented model is used to generate predictions. The model thus predicts that very prominent distractors can cause facilitatory interference in target-match conditions. This corresponds to observation (A3) in the literature review.

Another observation of the literature review was that the inhibitory effect in target-match conditions becomes easier to detect with increased distractor prominence (A2 vs. A1). However, below a base-level of 2, the original model shows no sensitivity of the inhibitory effect size in target-match conditions to distractor activation: The effect for a low activated distractor with a base-level activation of -0.5 is the same as for a highly activated distractor with a base-level value of 1. The reason is that Equation (2.3), repeated here as (3.1), which defines the activation spread from a cue to an item, is based on the *number* of competitors but not on their activation:

$$S_{ji} = S - \ln(fan_{ji}) \tag{3.1}$$

This means that the prediction of similarity-based interference in ACT-R (the *fan effect*) is not sensitive to the difference in activation between target and distractor.

For example, the standard model does not predict that a distractor in subject position in target-match conditions (15) causes more inhibitory interference than a distractor in object position (16).

(15) The tough soldier$^{+masc}_{+subj}$ that $\mathbf{Fred}^{+masc}_{+\boldsymbol{subj}}$/$\mathbf{Katie}^{-masc}_{+\boldsymbol{subj}}$ treated in the military hospital introduced himself$\{{}^{masc}_{subj}\}$ to all the nurses.

(16) The surgeon$^{+masc}_{+subj}$ who treated Jonathan$^{+masc}_{-subj}$/Jennifer$^{-masc}_{-subj}$ had pricked himself$\{{}^{masc}_{subj}\}$ with a used syringe needle.

Patil et al. (2012) found an inhibitory interference effect of the gender manipulation in target-match conditions as in (15), while Sturt (2003), using a design as in (16), did not find an effect. In an attempt to model the influence of the distractor position in ACT-R, Patil et al. used a *subject* feature that was possessed by both noun phrases and requested by the retrieval specification. In (15), the distractor would match the subject cue while in (16) it would not. Hence, the distractor would be more activated in (15) than in (16) relative to the target, which carries the subject feature in both designs. A higher distractor activation would then intuitively be expected to result in a stronger inhibitory interference effect in (15). However, in Patil et al.'s simulations, the target-match interference effect was *not* predicted to be stronger in (15) vs. (16). The reason was that, even when the distractor carries the requested subject feature and is thus activated more highly than without it, the distractor-match (interference) condition and the distractor-mismatch condition differ only in the match or mismatch of the gender cue. Since the fan effect depends on the number of cues matching in the interference condition compared with the no-interference condition, the standard ACT-R model predicts the same target-match interference effect for the designs of both Sturt (2003) and Patil et al. (2012). For the same reasons, a principled match-mismatch asymmetry, which was explained in Wagers et al. (2009) by the activation difference between target and distractor, is in fact not predicted in ACT-R.

We should, however, note that a principled asymmetry between target-match and target-mismatch effects can be introduced into the model by using specific parameter settings, namely by increasing the *maximum associative strength parameter* (MAS) (see Figure 3.6), which has a default of 1. A high MAS (S in the similarity-based interference equation 2.3) decreases the effect of similarity-based interference in general.[6] Hence, with a high MAS, similarity-based interference effects would only be predicted for large differences between conditions, i.e., a big fan effect, or for a less activated target. According to Equation (2.3), the fan is increased by the presence of more competitors. So, a high-MAS model would make the correct predictions for different target-match effects in, e.g., the authors' Experiments 1 and 2 in Jäger, Engelmann, and Vasishth (2015), where an effect was only found when multiple distractors were present. However, there are two concerns with this kind of model: Firstly, it relies on a specific relation between the MAS parameter and the absolute activation of the target. Secondly, just like the MAS=1 model, the predictions of this approach for inhibitory effects in target-match conditions are not sensitive to the difference in activation between target and distractor (see Figure 3.7, triangles). We therefore propose a superior approach in the next section.

3.2.6.3 Conclusion

The verbal predictions (A) and (B) do approximately match the computationally derived predictions of the LV05 model. When modeling distractor prominence with variable distractor

[6] A high *maximum associative strength parameter* (MAS) increases an item's activation such that the *fan* has only a relatively small effect. Since Equation (2.5), which maps activation to retrieval latency, is a negative exponential function, latency effects are diminished if the difference in activation *between two conditions* is relatively small compared to the item's absolute activation.

activation, the model can explain the absence of effects in target-mismatch conditions (B1) and facilitatory interference in target-match conditions (A3). However, the size of similarity-based interference is not sensitive to the difference in activation between target and distractor. Hence, the model does not predict increasing inhibitory effects in target-match conditions as a result of increasing distractor prominence (A1 vs. A2). Furthermore, the original model is unable to explain inhibitory effects in target-mismatch conditions (B3).

We propose two principles as an extension to an ACT-R cue-based retrieval model: The *prominence correction* predicts a relation between distractor prominence and inhibitory effect sizes in target-match conditions, and the concept of *associative cues* predicts that in certain linguistic environments cues can be confused, which leads to inhibitory interference in target-mismatch conditions. The next section introduces our extended model of cue-based retrieval and explains the proposed principles in detail.

3.3 An Extended Cue-Based Retrieval Model

Table 3.2 in the introduction summarizes the mechanisms of the extended model and how the observations (A1-3) and (B1-3) are explained in the model. The *prominence correction* is a general mechanism that makes similarity-based interference sensitive to the activation difference between target and distractor. It thus predicts small or missing effects in target-match conditions with low distractor prominence (A1) and inhibitory effects for increased distractor prominence (A2). The principle of *associative cues* introduces the possibility of similarity-based interference in target-mismatch conditions. This predicts inhibitory interference in target-mismatch conditions in certain linguistic environments where retrieval cues are confused (B3).

3.3.1 Principle 1: Prominence Correction

It is intuitive to assume that a perfectly matching antecedent is less affected by interference than a partially matching one. In addition, a highly activated distractor should increase the interference effect. In the literature review, we observed that 1) the probability of finding an effect may increase with distractor prominence such as grammatical position and discourse saliency of the distractor and 2) that cases of facilitatory interference in target-match conditions are connected with specifically high distractor prominence.

These observations are predicted when we assume that distractor prominence is represented by the distractor's activation in memory and that the strength of similarity-based interference is modulated by the relative activation of the competing items. We, however, also showed that the SBI component of in ACT-R (the *fan effect*) is not sensitive to distractor activation. Hence, the original cue-based retrieval model does not predict 1) increased effects by prominence, while, on the other hand, it does predict 2) facilitatory interference for high prominence due to misretrievals.

We extended ACT-R with a *prominence correction* which scales the strength of the *fan effect* (the inhibiting effect of similarity-based interference) by the activation difference between target and distractor.

3.3.1.1 Implementation

The *prominence correction* is the left part of the product in Equation (3.2). It transforms the fan_{ji} for cue j of item i using a logistic function of the difference in activation between the item and its competitors:

$$fan'_{ji} = \begin{cases} \frac{1}{1+\exp\{-C(x_0-[A_i-\bar{A}_{Comp}])\}} \times fan_{ji}, & \text{if } C > 0 \\ fan_{ji}, & \text{otherwise} \end{cases} \tag{3.2}$$

where $[A_i - \bar{A}_{Comp}]$ is the difference between the activation A_i of item i and the mean activation $\bar{A}_{Comp}$ of all competitors associated with cue j. C is the *prominence correction factor* (PCF), which scales the strength of the prominence effect on the fan. The value of x_0 is the *prominence correction offset.*

As Figure 3.8 illustrates, the prominence correction is a logistic function of the activation difference between target and competitors $[A_i - \bar{A}_{Comp}]$ which multiplies the original fan_{ji} with a value between 0 and 1. For highly active competitors, the activation difference is negative and the prominence correction approaches 1, which means that the resulting fan' equals the original *fan.* Hence, for very highly activated distractors, the extended model and default ACT-R make the same predictions (see also Figure 3.9). On the other hand, if the target is more highly activated than the competitors and $[A_i - \bar{A}_{Comp}]$ is positive, the correction function decreases and approaches 0. This decreases the second term in Equation (2.3) with the consequence that the target's activation is less affected by its competitors, diminishing similarity-based interference effects. The *prominence correction factor* C scales the steepness of the logistic function and was set to 5 for all simulations of the extended model presented here. The *offset* x_0 defines the curve's midpoint, meaning that $fan'_{ji} = 0.5 \times fan_{ji}$ at an activation difference between target and distractor of x_0. It was fixed at 1.3 for all simulations.

3.3.1.2 Predictions

The extended model predicts by the same mechanism the absence and presence of target-match interference effects and occasional observations of facilitation. Figure 3.9 compares the predictions of our extended model (triangles) with the predictions of the original model (circles)

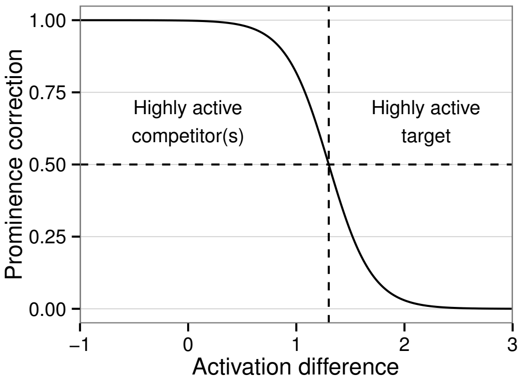

Figure 3.8: Prominence correction by activation difference $[A_i - \bar{A}_{Comp}]$ with $C = 5$ and $x_0 = 1.3$.

and the data means (big, gray circles) for three prominence classes (no prominence, subject *or* topic, subject *and* topic). The extended model makes the following predictions with respect to *target-match* conditions: (A1) With a distractor base-level activation equal to the target base-level activation (at 0), the predicted effect is close to zero; (A2) inhibitory effects increase with increasing distractor prominence; (A3) at very high distractor prominence, the effect becomes facilitatory due to an increased number of misretrievals. Numerically, the predictions of the extended model (using standard parameters) very closely match the data means.

Thus, distractor prominence could explain qualitative differences in target-match effects, e.g., between (A1) Badecker and Straub (2002), Experiment 5 (no effect), (A2) Badecker and Straub (2002), Experiment 3 (inhibition), and (A3) Cunnings and Felser (2013), Experiment 2 (facilitation) by differences in distractor activation due to the distractor being (A1) inside a genitive phrase, (A2) in subject position, or (A3) in topicalized subject position, respectively.

Controlled experiments for testing prediction (A2) were conducted by Cunnings and Felser (2013) and Patil et al. (2012). Both demonstrated an effect of distractor position by using the sentence material of Sturt (2003)'s Experiment 2 and moved the distractor into subject position. Patil et al. found an inhibitory interference effect of the gender manipulation in target-match conditions, which had not been found in Sturt's original Experiment 2. As discussed in the previous section, the standard ACT-R model used by Patil et al. did not predict a stronger inhibitory interference effect in target-match conditions for a distractor in subject position, because similarity-based interference in this model is not sensitive to the difference in activation between target and distractor. However, as Figure 3.9 shows, with the prominence correction, our extended model is sensitive to distractor prominence and predicts a larger inhibitory interference effect (A2) in designs using a subject distractor (Patil et al., 2012) than in designs using an object distractor (Sturt, 2003, Experiment 2).

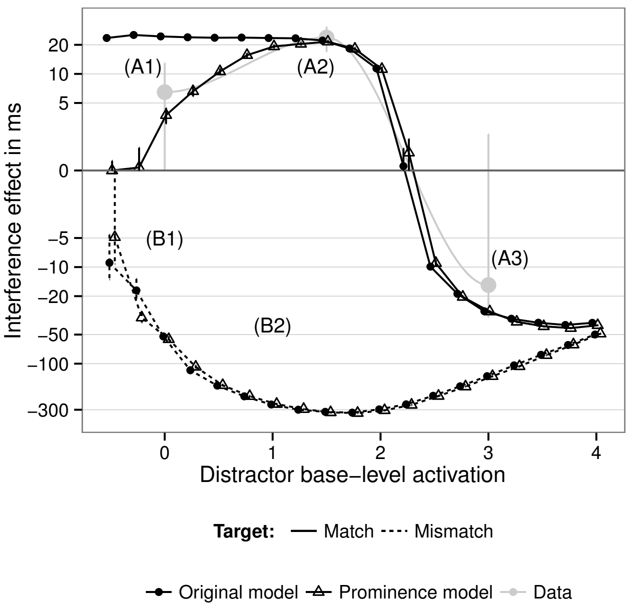

Figure 3.9: Relation between the distractor base-level activation B_{Distr} and the predicted interference effect (logarithmic scale) when target base-level is zero. Three models are compared: Original ACT-R with *prominence factor* $C = 0$ (circles) and the *prominence model* with $C = 5$ (triangles). Solid lines represent the conditions where the target matches all retrieval cues; dashed lines represent conditions where the target is only a partial match. The light gray line represents the mean of empirical target-match effects by distractor prominence as a smoothed curve connecting the data means in the three prominence classes as shown in Figure 3.5. Error bars represent standard errors.

In contrast to Patil et al., Cunnings and Felser (2013) found a *facilitatory* effect (for low-span readers) in target-match conditions in their experiment using a distractor in subject position. The difference to Patil et al.'s experiment was that, in Cunnings and Felser's design, the distractor was not just in subject position but also topicalized. The same is true for Sturt (2003)'s Experiment 1, which also used a topicalized subject distractor, although in a proactive design, and found facilitatory interference.

The combination of the distractor being topicalized and being in subject position can be assumed to make it very salient to the reader. The results of Sturt (2003) and Cunnings and Felser (2013) support this assumption. In our extended model, a high saliency would increase the distractor's base-level activation, which predicts the observed facilitatory effect, as is illustrated in Figure 3.9.

3.3.2 Principle 2: Associative Cues

Similarity-based interference (SBI) is the mechanism that slows down retrieval. Occasional misretrievals can, on the other hand, lead to a speed-up in mean retrieval latencies. In the standard ACT-R model, SBI is only present in *target-match* conditions, which excludes the prediction of inhibitory interference in *target-mismatch* conditions. As a consequence, there is no principled way to explain observation (B3), inhibitory interference effects in target-mismatch conditions, within the standard ACT-R-based LV05 model of cue-based retrieval.

To illustrate this point again, recall the reflexives example (10), repeated here as (17).

(17) a. *Target-match; distractor-match*
The surgeon$^{+masc}_{+\,c\text{-}com}$ who treated Jonathan$^{+masc}_{-\,c\text{-}com}$ had pricked himself$\{{}^{masc}_{c\text{-}com}\}$

b. *Target-match; distractor-mismatch*
The surgeon$^{+masc}_{+\,c\text{-}com}$ who treated Jennifer$^{-masc}_{-\,c\text{-}com}$ had pricked himself$\{{}^{masc}_{c\text{-}com}\}$

c. *Target-mismatch; distractor-match*
The surgeon$^{-fem}_{+\,c\text{-}com}$ who treated Jennifer$^{+fem}_{-\,c\text{-}com}$ had pricked herself$\{{}^{fem}_{c\text{-}com}\}$

d. *Target-mismatch; distractor-mismatch*
The surgeon$^{-fem}_{+\,c\text{-}com}$ who treated Jonathan$^{-fem}_{-\,c\text{-}com}$ had pricked herself$\{{}^{fem}_{c\text{-}com}\}$

In target-match conditions (17a) vs. (17b), similarity-based interference arises because, in the interference condition (a), the target *surgeon*$^{+masc}_{+\,c\text{-}com}$ and the distractor *Jonathan*$^{+masc}_{-\,c\text{-}com}$ share the manipulated *masculine* feature that is requested by the reflexive *himself.* In ACT-R terms, both items enter into a competition for a limited amount of activation that is spread to every feature matching the retrieval specification. As a consequence, both items' activation is reduced by the *fan effect* of Equation (2.3), predicting an inhibitory interference effect. In target-mismatch conditions (17c) vs. (17d), on the other hand, no similarity-based interference arises because *surgeon*$^{-fem}_{+c\text{-}com}$ and *Jennifer*$^{+fem}_{-c\text{-}com}$ do not overlap in the manipulated *feminine* feature. Similarity-based interference arises when the experimentally manipulated feature is requested by the retrieval specification *and shared between target and distractor.* This leads to the standard ACT-R prediction of cue-based retrieval that inhibitory interference effects are only predicted in *target-match* conditions but not in *target-mismatch* conditions.

We argue that this prediction rests on the simplified assumption of a categorical *yes-or-no* match between cues and features. As an alternative, we propose the concept of *associative cues*: Instead of an absolute match/mismatch relation, cues and features are related with a continuously valued associative strength such that a cue can be associated with multiple features to different degrees. The stronger the association between a cue and a feature, the more efficient the cue is for activating items with that feature in memory. In the context of the ACT-R architecture underlying our model, an associative relation between cues and features, though not implemented, is a rather straightforward assumption. In standard ACT-R, any two memory items can be assigned a mutual similarity. This includes feature values and cue values as these are items in memory, too. Similarities are used, for example, in Equation (2.4) to enable the retrieval of items that do not match the retrieval cues but might nevertheless be similar (like an orange item when cueing for *red*). There is no reason not to use the same similarities in the calculation of the fan effect of Equation (2.3), since it is only natural to assume that feature similarity should not only affect retrieval probability but also similarity-based interference.

Theoretically, we motivate the concept of associative cues by the assumption of a learning mechanism that is based on the frequency of occurrence: From the perspective of learning, the relevance of individual features for the completion of a certain dependency is a heuristic shaped by experience. If two features co-occur frequently in target items for a certain type of dependency, it does not matter for the success of completing that dependency whether these two features can be distinguished conceptually.

As an example, consider the difference between the linguistic environments of reflexives and of reciprocals: English reflexives exhibit several forms like *himself*, *herself*, *itself*, and *themselves*. Thus, a correct reflexive antecedent can appear in a number of different combinations of the structural feature with number and gender: $\{\substack{+\mathit{fem/masc/neut},\ +\mathit{plur/sing} \\ +\mathit{c\text{-}com}}\}$. Clearly, a dissociation of these features from each other in the retrieval request is beneficial for identifying the correct target with respect to the individual form of the reflexive. On the other hand, the correct target for a reciprocal invariably requires the feature pair $\{\substack{+\mathit{plural} \\ +\mathit{c\text{-}com}}\}$. Every time a c-commanding item is required by a reciprocal, the correct one will also be plural, and vice versa. Because +*plural* and +*c-com* always co-occur, learning a successful retrieval specification for this specific environment does not require a strong dissociation between these two features. Instead, it can be thought as more efficient to also activate plural items with the *c-com* cue and vice versa. Alternatively, one could also think of it as a resource-preserving strategy that avoids unnecessary effort of a precise one-to-one mapping from a cue to a feature. In any case, the cues *c-com* and *plural* would both be associated to some degree with the features +*c-com* and +*plural.* For example, the *associative strength* of the *c-com* cue could be 100% with the feature +*c-com* and 25% with the feature +*plural.* This means that at retrieval, a +*plural* feature would receive 25% of the amount of activation a +*c-com* feature receives. Similarly, the associative strength of the *plural* cue would be 100% with the feature +*plural* and 25% with the feature +*c-com.* We call this a *crossed association* between the cues *c-com* and *plural.* In other words, the two cues are being *confused.*

Figure 3.10 illustrates how a confusion of *c-com* and *plural* on a level of 25% leads to inhibitory interference in target-mismatch conditions. Panel (1) shows target-mismatch conditions under the standard assumption of a one-to-one association between cues and features with no cue confusion. The target (with a +*c-com* feature) receives the same amount of spreading activation no matter whether the distractor matches the *plural* cue (b) or not (a). Consequently, no inhibitory interference arises. Panel (2) shows the same conditions under the assumption of cue confusion. Here, each cue is associated with 25% strength with the feature that is matched by the other cue. This means for the distractor-mismatch condition (2a) that the target, which carries the +*c-com* feature, receives 100% of the available activation from the *c-com* cue and another 25% from the *plural* cue. In the distractor-match condition (2b), now the distractor carries the +*plural* feature. As a consequence, the activation that is spread from the *c-com* cue has to be divided between target and distractor. Since the associative strength between *c-com* and +*plural* is 25%, it receives 25% of the amount of activation the *c-com* receives. This corresponds to a ratio of 20/80. The target can no longer receive 100% of the activation from the *c-com* cue but has to give away 20%. In the same way, the distractor does not receive the maximum amount of activation spread from the *plural* cue because some of it goes to the target. Thus, in the distractor-match condition (2b), target and distractor are reduced in their activation by the same mechanism of similarity-based interference which is assumed for target-match conditions. Consequently, inhibitory interference is predicted.

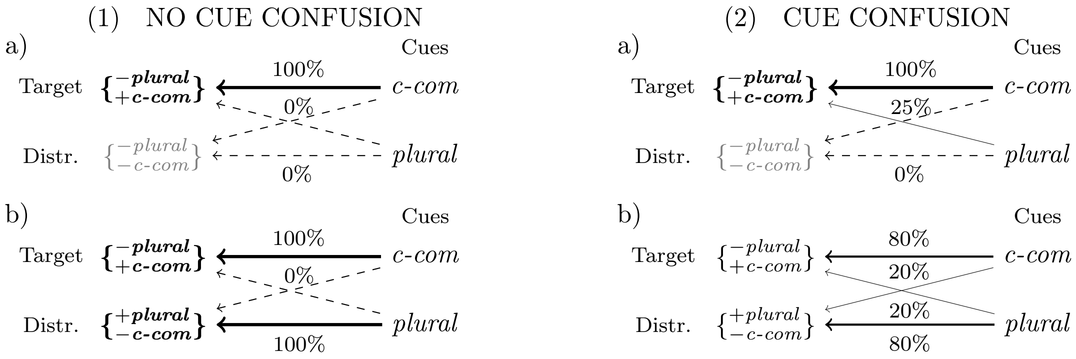

Figure 3.10: Spreading activation from cues to items in target-mismatch conditions under standard assumptions (1) and under cue confusion (2). Distractor-mismatch (a) and distractor-match (b) conditions are shown. Each condition has two memory items as feature sets on the left and retrieval cues on the right. Connecting lines and numbers represent the amount of activation spread from a cue to an item as percentage of total available activation per cue. Left panel (1) shows classical one-to-one cue-feature associations (e.g., English reflexives); right panel (2) shows crossed associations (e.g., reciprocals). Inhibitory interference is predicted in (2b) vs. (2a).

3.3.2.1 Implementation

In order to take into account feature similarity, we adjusted the *fan* computation by reusing the similarity measure M from (2.4). The original *fan* equation simply sums over items that absolutely match the retrieval cue in question. Our new *fan* equation in (3.3) sums the similarities of cue j with all features k that are part of other items in memory:

$$fan_{ji} = 1 + \sum_{k}(1 + M_{jk}) \tag{3.3}$$

where M_{jk} is the similarity of cue value j with feature value k on a scale of $[-1, 0]$, with 0 meaning identity and and -1 being the maximum difference. The similarity value M_{jk} here corresponds to the *associative strength* between cue and feature. This associative strength is assumed to be dynamically adaptive between individual linguistic environments, meaning that the value M_{jk} changes based on experience in different dependencies or languages. Equation (3.3) predicts that the stronger a cue-feature association M_{jk} the more this feature will contribute to similarity-based interference related to that cue. If, for instance, the association $M_{\textit{c-com;plur}}$ for reciprocals is -0.75, the resulting fan for the *c-com* cue would be 1.25 instead of 1.

Below, we will, for reasons of simplicity, always refer to the confusion between two specific cues as *cue confusion level* CL, which is treated as a parameter that can be set for a pair of cues in different linguistic contexts. For example, a cue confusion level of 25% like in the reciprocals example above corresponds to the case where $M_{\textit{c-com;+plur}} = M_{\textit{plur;+c-com}} = -0.75$ as of Equation (3.4), whereas $M_{\textit{c-com;+c-com}} = M_{\textit{plur;+plur}} = 0$.

$$M_{\textit{cue1;feature2}} = M_{\textit{cue2;feature1}} = \mathrm{CL}/100 - 1 \tag{3.4}$$

As demonstrated above, at a confusion level of 25%, the available activation is divided between two features at the ratio 20/80. A level of 0% corresponds to 0/100, which is the classical case of maximal difference between non-identical cue and feature values: $M_{jk} = -1$. A level of 100% corresponds to maximal confusion between two cues, as if the values were identical with each other's features: $M_{jk} = 0$. In this case, an equal amount of activation (a ratio of 50/50) spreads to all associated features.

3.3.2.2 Predictions

Figure 3.11 illustrates the effect of cue confusion for a usual target-mismatch design. With an increasing level of cue confusion the facilitatory effect in target-mismatch conditions decreases and turns into an increasing inhibitory effect, predicting observation (B3).

As explained above, our theory of co-occurrence-induced cue confusion predicts a higher confusion level for reciprocals than for English reflexives. This could explain the result of Kush and Phillips (2014), who found inhibitory interference in target-mismatch conditions in Hindi reciprocals. However, the level of cue confusion might also vary within the same dependency type across languages. An example potentially involving cue confusion is the Mandarin reflexive *ziji*. In Experiment 1 in Jäger, Engelmann, and Vasishth (2015), we found an inhibitory interference effect in target-mismatch conditions. This can be explained by our proposal in a way similar to the case of reciprocals: As there is no gender or number information in the Mandarin reflexive, the correct antecedent is invariably required to be c-commanding and animate. Hence, under the cue confusion account, the confusion level compared to English reflexives is predicted to be higher for Mandarin *ziji*, just like for reciprocals.

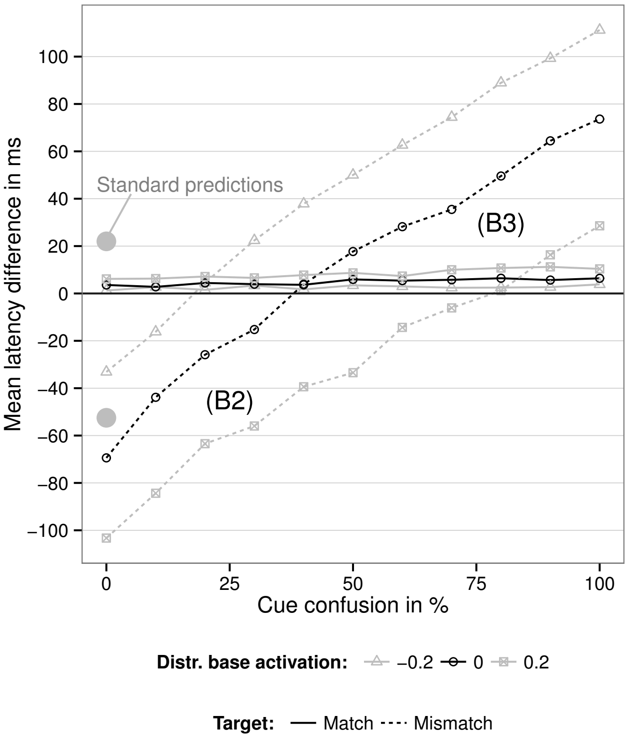

Figure 3.11: Predicted interference effects by cue confusion level for the extended model with distractor base-level activation of −0.2 (triangles), 0 (circles), and 0.2 (squares). The standard predictions of original ACT-R without cue confusion or distractor prominence are represented by the gray points at cue confusion of 0%.

Under a theory of cue confusion, an interesting question is whether categorically distinguishing two cues requires cognitive effort. If so, one would expect an additional variation of the confusion level that depends on task demands and individual differences. There is evidence that the depth of linguistic processing is influenced by task-specification (Logačev & Vasishth, 2015; Swets et al., 2008) and individual differences (von der Malsburg & Vasishth, 2013; Traxler, 2007), resulting in underspecification of sentence representations or “good-enough processing” (Ferreira et al., 2002). In the same way, cue confusion could be part of a dynamically adapted resource-preserving strategy. This question can only be answered by further experiments. However, there is one experiment in our review that tested participants on working memory capacity: Cunnings and Felser (2013) found in their Experiment 2 on English reflexives an inhibitory effect on the critical region in target-mismatch conditions only for low-capacity readers. The effect is only marginally significant but would be in line with the assumption of an individual-level variation of cue confusion due to adaptive processes. This assumption predicts elevated confusion levels for readers with less cognitive resources in order to preserve speed. It also predicts increased cue confusion for experiments with little task demand, like easy comprehension questions, because the effort of a precise cue specification would not be necessary.

In summary, the *principle of associative cues* predicts that cues can be associated with several features to different degrees depending on experience with the linguistic context. Crossed cue-feature associations between two cues lead to *cue confusion* and predict decreased facilitation or even inhibitory interference in *target-mismatch* conditions (B3) for dependency environments with high *feature-co-occurrence* in comparison to environments with less feature-co-occurrence. Furthermore, it is possible that also task-demands and individual differences influence the level of cue confusion.

3.4 Simulations

In order to thoroughly test the predictions of the standard LV05 cue-based retrieval model and of our extended model, we simulated all experiments listed in our review that used either eye tracking or self-paced reading as experimental method. These methods reflect processing times that can be compared to retrieval latencies predicted by the model. Studies on agreement attraction were not included in the simulations because, as discussed above, the results in target-match conditions consistently contradict the basic predictions of cue-based retrieval and thus point to an underlying mechanism that is different from other subject-verb dependencies and reflexive anaphora. Simulations were conducted with the standard model and with the extended model that implemented *prominence correction* and *associative cues.*

The presented simulations can be run online using an interactive app created with Shiny by RStudio (RStudio & Inc., 2014) at the address `https://engelmann.shinyapps.io/inter-act`.

3.4.1 Method

Simulations were carried out with 2000 iterations per study with each the original and the extended model. Initial parameters were set to ACT-R defaults or values used in previous simulations (e.g., Lewis & Vasishth, 2005): *latency factor* LF $= 0.15$ *activation noise* ANS $= 0.15$, *mismatch penalty* MP $= 1.5$, *retrieval threshold* RT $= -1.5$. The extended model introduces three new parameters: The *prominence correction factor* PCF, *prominence correction offset* x_0, and *cue confusion level* CL. Setting PCF and CL to zero corresponds to the classical LV05 model. For all simulations with the extended model, PCF was set to 5, and the offset to 1.3. The confusion level CL was set to 0% by default and was estimated for the two environments discussed above, reciprocals and Mandarin reflexives, which exhibit increased feature-co-occurrence. Finally, *distractor base-level activation* was estimated for each study individually. The resulting values are plotted in Figure 3.13.

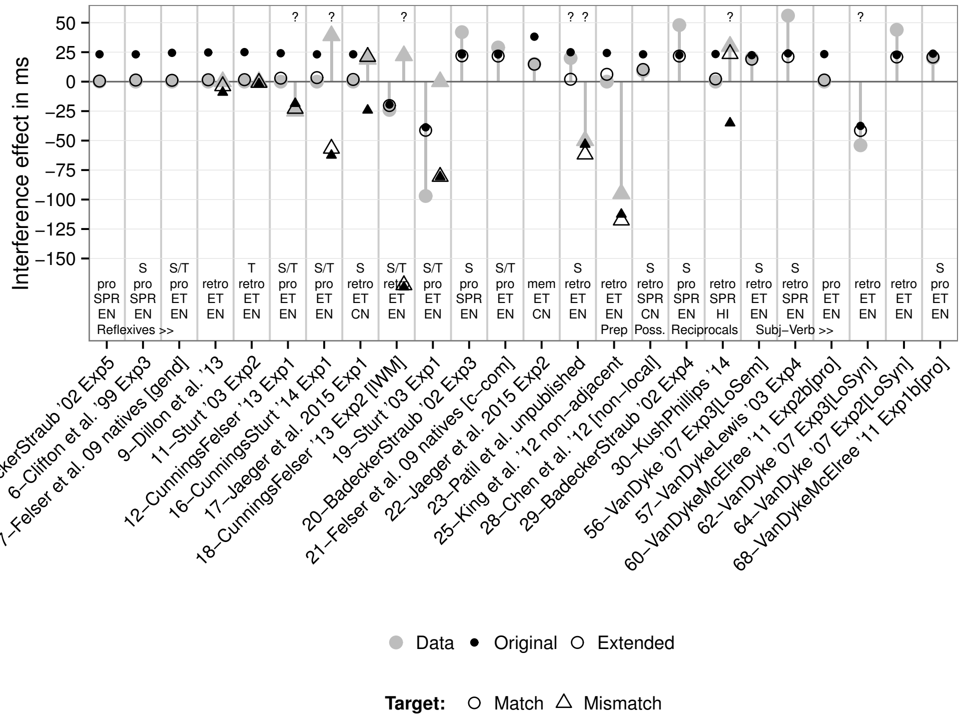

Figure 3.12: Data and simulation results of interference effects in target-match (circles) and target-mismatch conditions (triangles) for selected studies. Data is shown in gray and includes bars for better visibility of the direction of the effect. The original model is shown as closed black symbols and the extended model as open black symbols. Effects in target-mismatch conditions are only shown where empirical data exists.

3.4.2 Results and Discussion

3.4.2.1 General Fit

Figure 3.12 shows the predictions for interference in target-match (circles) and target-mismatch conditions (triangles) of the extended model (open symbols) and the original ACT-R model (black filled symbols) in comparison with the data (gray filled symbols). For reasons of space, the figure shows only selected studies. The selection is representative in the sense that it reflects all match/mismatch effect patterns and combinations of language, method, distractor prominence, and interference type found in the review. If, e.g., there were two similar experiments with a similar outcome (e.g., Badecker & Straub, 2002, Experiments 5 and 6), we report only one of them, since model fit and parameters would be equal. At the bottom of the plot, there are five

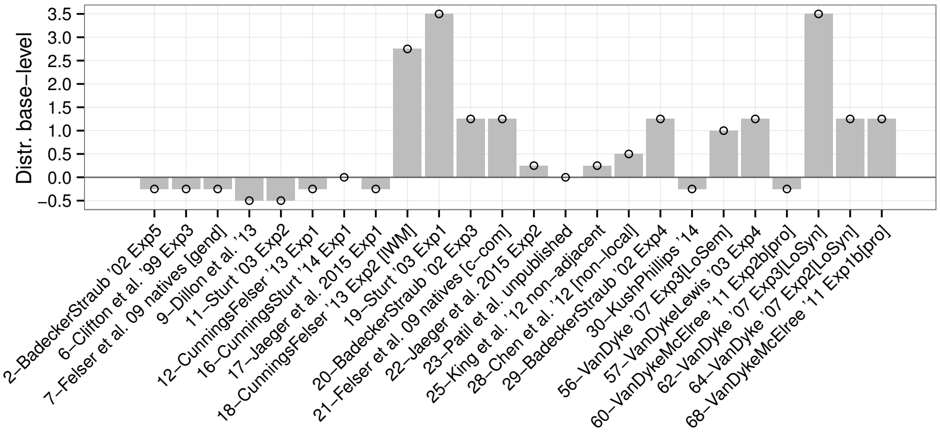

Figure 3.13: Estimated values for distractor base-level activation.

lines that indicate the following experimental characteristics: Distractor prominence (subject, topic, or both), interference type (proactive, retroactive, or memory task), method (SPR or eye tracking), language, and the type of dependency. On the top of the graph, question marks '?' indicate where an effect was only marginally significant.

Estimated distractor base-level activation values are shown in Figure 3.13. The cue confusion values for reciprocals and Mandarin reflexives were estimated at 30% and 25%, respectively. By fitting distractor base-level activation, both the original and the extended model predict facilitatory interference in target-match conditions for high distractor prominence (A3). However, due to the *prominence correction*, the extended model (open symbols in Figure 3.12) accounts for more variation in target-match effect sizes, and *cue confusion* enables it to predict inhibitory interference in target-mismatch conditions in two cases. The root-mean-squared error RMSE across all simulations (also those not shown in the plot) was 42.72 for the original model (match: 25.05; mismatch: 75.07) and 37.81 for the extended model (match: 18.1; mismatch: 70.25). The correlation was 0.38 for the original model (match: 0.69; mismatch: 0.14) and 0.5 for the extended model (match: 0.91; mismatch: 0.31). The extended model numerically fits the data better as it captures more variation. The most significant improvement is, however, in the correlation of target-match effects, which is due to the prominence correction making similarity-based interference in target-match conditions sensitive to distractor prominence. In the following, we analyze the predictions in detail and discussed the meaning of the estimated values.

3.4.2.2 Distractor Prominence

We analyzed the relation between distractor prominence and estimated distractor base-level activation within the same three prominence classes used in the literature review: (0) neither subject position nor topicalized, (1) either subject position or topicalized, and (2) both subject position *and* topicalized. The means of the estimated distractor base-level activation for the three classes were 0.33, 0.54, 1.17, respectively. The correlation between prominence and distractor base-level activation across all simulations was only 0.26. However, it seems that distractor activation was estimated generally higher (mean = 1.02) for subject-verb dependencies (Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006, 2011) compared to reflexives (mean = 0.32). Looking only at studies on reflexives, the correlation between prominence and distractor base-level activation was 0.51, with prominence class means of −0.17, 0.22, 1.17. A multiple linear regression model on all simulations studies confirmed a significant effect of our hypothesized distractor prominence measure (coded as an ordinal factor) on estimated distractor base-level activation (estimate = 1.02, SE = 0.47, t = 2.17, p = 0.04). Dependency type was marginally significant, with subject-verb dependencies predicting higher activation (estimate = 0.94, SE = 0.47, t = 2, p = 0.06). The other factors language, method, and interference type were not significant ($p > 0.6$).

The account of distractor prominence does, however, not explain all variation. There are three studies with the distractor in topicalized subject position that show zero target-match effects: Felser et al. (2009) (gender, ID 7), Cunnings and Felser (2013), Experiment 1 (ID 12), Cunnings and Sturt (2014), Experiment 1 (ID 16). It is possible that these missing effects represent type II errors. A way to reconcile these missing effects with the prominence assumption would be to estimate the base-level activation in these cases not at zero but just in the range where in Figure 3.7 the predicted target-match effect size crosses the zero line before turning into a facilitation, which is between a base-level of 2 and 2.5. If we assumed this is the case and set distractor base-level for these experiments to 2.25, the correlation between prominence and distractor base-level activation for studies on reflexives would increase from 0.51 to 0.81.

Other studies had the distractor either in subject or in topicalized position and did not find a target-match effect (Chen et al., 2012; Clifton et al., 1999; Jäger, Engelmann, & Vasishth, 2015; Parker & Phillips, 2014; Sturt, 2003; Van Dyke, 2007). This suggests that subject position and topicalization are factors that slightly increase the probability of finding an effect, whereas the combination of both has more reliable consequences.

As pointed out above, distractor activation was estimated higher in general for subject-verb dependencies, independent of the distractor prominence status. Almost all subject-verb studies included in the simulations show large effect sizes, which might point to a high saliency of the distractor in these experimental designs. The only exception is Van Dyke and McElree (2011), Experiment 2 (IDs 60 and 61), where no effects were found. Here, the distractor base-level activation was estimated in the range comparable to reflexive studies. This is plausible when interpreting this experiment in the context of Experiment 1 reported in the same publication: Van Dyke and McElree's Experiment 1 (IDs 68 and 69) is different from Experiment 2 in that it had the distractor in subject position while it was in object position in Experiment 2. Fitting with the prominence account, an inhibitory interference effect in target-match conditions was found in Experiment 1 but not in Experiment 2.

The large effects for subject-verb studies manipulating structural cues like in Van Dyke and Lewis (2003) (ID 57) and Van Dyke (2007) (IDs 55 and 56) can be explained by the experimental design: These experiments manipulated distractor position as being the subject of a complement clause vs. being contained in a prepositional phrase, in this way manipulating cue match and prominence at the same time. A large target-match effect is also expected in Van Dyke and McElree (2006) (ID 65) because it required the participants to rehearse three distractors in memory while reading the sentences. Only in the semantic cue-manipulation versions of Experiments 1 and 2 of Van Dyke (2007) (IDs 63, 64), the large inhibitory effects cannot be straightforwardly explained by a distractor position. In both experiments, the distractor was a verb-modifying prepositional phrase inside a relative clause. However, the distractor is in both cases directly adjacent to the critical verb. This could cause strong inhibitory interference due to local coherence or simply the large recency difference between target and distractor.

In concordance with the prominence assumption, the facilitatory interference effects in target-match conditions (A3) observed in Cunnings and Felser (2013), Experiment 2 (ID 18) for low-span readers and in Sturt (2003), Experiment 1 (ID 19), can be explained by high prominence (distractor in topicalized subject position). On the other hand, the facilitatory effect in Van Dyke (2007), Experiment 3 (LoSyn, ID 62) is not connected to increased distractor prominence such as structural position or discourse saliency. Note, however, that the facilitatory effect in this experiment was only significant in the by-participants analysis in regression-path duration on the post-critical region.

3.4.2.3 Cue Confusion

There are four cases of inhibitory interference in target-mismatch conditions among the simulated experiments: Cunnings and Sturt (2014), Jäger, Engelmann, and Vasishth (2015), Cunnings and Felser (2013), and Kush and Phillips (2014). We consider all four effects although, except for Jäger, Engelmann, and Vasishth (2015), they were only marginally significant. The cue confusion level was estimated for reciprocals and Mandarin reflexives at 30% and 25%, respectively, which correctly predicts the results of Kush and Phillips (2014) (ID 30) and Jäger, Engelmann, and Vasishth (2015), Experiment 1 (ID 17). For the inhibitory effect in Cunnings and Sturt (2014), Experiment 1 (ID 16), we have no explanation whatsoever under our proposed account. In Cunnings and Felser (2013) (ID 18), an inhibitory effect was found for readers with low memory span. It may be possible that low-span readers generally experience an increased confusion level, which could explain this effect. However, in the current model an inhibition in target-mismatch conditions due to cue confusion is implausible when at the same time the distractor is so highly activated that it causes facilitation in target-match conditions. This is illustrated in Fig 3.11: Already at a distractor base-level activation of 0.2, inhibitory interference in target-mismatch is only predicted for a cue confusion level close to 100%.

3.4.2.4 Cue Strength

As discussed in the literature review, a study on reflexives by Parker and Phillips (2014) presented evidence in support for cue-weighting. They found large significant facilitatory effects in conditions where the target was mismatched in *two* features with the retrieval cues, while there was no effect in the usual target-mismatch conditions when the target was mismatched with only one feature. Parker and Phillips concluded that cues must be weighted and used simulations to show that structural cues have to be weighted three times as strong as semantic cues.

We have simulated the six conditions of their experimental design with our extended model and found that it is not necessary to assume differently weighted cues in order to obtain their result. With a distractor base-level activation of 0.2, our model predicts very small effects in target-match and 1-feature mismatch conditions, and a huge facilitatory effect in 2-feature mismatch conditions (see Table 3.7). The slightly elevated distractor base-level of 0.2 is justified by the fact that the distractor in Parker and Phillips's design was in subject position and, thus, prominent.

Table 3.7: Parker and Phillips (2014), Experiment 1 (TFT) data vsp̀redictions of the extended model with distractor base-level activation of 0.2.

Target	**Data**	**Model**
Match	n.s.	6 ms
1-Mismatch	n.s.	−2 ms
2-Mismatch	−250 ms	−326 ms

We conclude that, at least in this case, the assumption of differently weighted cues may not be necessary. Instead the correct predictions can be derived from the more general principle of the prominence correction. We have not included a manipulation of cue strength in the simulations presented here. However, we have implemented a cue-weighting mechanism in the extended model, which can be used for further simulations with our online application at `https://engelmann.shinyapps.io/inter-act`.

3.4.2.5 Summary

The extended cue-based retrieval model fits the data better than the original model; the improvement is largely due to the prominence correction leading to a better fit in target-match conditions. Distractor prominence is significantly correlated with base-level activation, especially in reflexive-antecedent dependencies. The model offers an explanation for the absence of effects in a number of studies (A1, B1) and the presence of effects in other studies (A2, B2), especially with regard to target-match conditions. It predicts increased effect sizes and an increased probability of finding an effect due to structural position or discourse saliency of the distractor. The combination of both structural position and discourse saliency can explain two of three cases of *facilitatory interference in target-match conditions* (A3). The assumption of cue confusion in the processing of reciprocals and Mandarin reflexives explains *inhibitory interference in target-mismatch conditions* (B3) in Kush and Phillips (2014) and Jäger, Engelmann, and Vasishth (2015), but the model cannot account for the marginal effects in Cunnings and Sturt (2014) and Cunnings and Felser (2013).

3.5 General Discussion

We have presented a comprehensive review of the literature on retrieval interference in subject-verb and reflexive-antecedent dependencies and discussed it with respect to cue-based retrieval theory. We have observed that the results of studies examining interference effects in number agreement, in particular the results in target-match conditions, are incompatible with the predictions of cue-based retrieval, or at least additional assumptions specific to this dependency are needed in order to explain the results. For non-number subject-verb dependencies and reflexive-antecedent dependencies, we have identified the two determinants distractor prominence and cue confusion as explanatory factors for a number of results that are not predicted by the standard ACT-R-based cue-based retrieval theory. An extended model of cue-based retrieval has been presented and evaluated with quantitative simulations of a large set of the reviewed experiments.

Our extended model offers an approach of quantifying the predicted influence of distractor prominence and cue confusion on retrieval interference. It thus enables a more informed decision on how to treat the results in question, i.e., which of them are more probably noise and which could be worthy of further investigation. A model with clear predictions on distractor prominence and cue confusion makes it possible to design controlled experiments that gain more insight into the relation between these factors and interference. There is no doubt that, in general, experiment-specific characteristics can influence the results, and it is hard to control for all possible influences. In order to meaningfully interpret a result in the context of previous literature, the question that needs to be answered is: What are the possible influences of factors that differ between the current and previous studies? Examples we have discussed here are language, experimental method, distractor position, and linear order of target and distractor. Another factor is the differences in the cue/feature contexts between reflexives and reciprocals. There may of course be more factors that have not yet been identified.

In the following, we discuss our findings and their implications.

3.5.1 Distractor Prominence

Our review suggests that there are various facets of distractor prominence that influence the dependencies under study in different ways. While the chance of finding interference effects in reflexive-antecedent and non-number subject-verb dependencies is apparently related to the grammatical role, the most important predictor for finding an effect in target-match conditions of number agreement is the linear order of target and distractor. The question here is whether these differences are due to systematic characteristics of the dependencies under discussion or whether they are just a result of the specific experimental designs used.

The strong relation between linear order and the proportion of effects in number agreement might be connected to the fact that most effects are found when the distractor is inside a prepositional phrase that builds a complex noun phrase with the target. This configuration introduces a close structural relation between the distractor and the target, which might already affect the encoding of the NPs, as the feature percolation account (Bock & Eberhard, 1993) states. Another potential problem that these constructions pose to an interpretation of the results is that, in a number of studies, the distractor is adjacent to the verb. As pointed out by Wagers et al. (2009), the observed effect on the verb could in these cases be a spillover effect from the distractor. As plurals are considered more complex and thus induce lexical processing difficulty (Lau et al., 2007; Lehtonen et al., 2006; New et al., 2004), a spillover effect from the distractor would make the plural condition harder. This could explain the observed facilitatory effects in singular-verb target-match conditions. In contrast, a potential spillover effect would not explain facilitatory effects in grammatical conditions with a *plural* verb, since there the plural-distractor condition is easier. However, in plural-verb grammatical conditions of number agreement, the results are inconclusive: Out of seven studies, there are only two that found a facilitatory effect (Acuña–Fariña et al., 2014 and Pearlmutter, 2000, Experiment 2) and one that found inhibition (Pearlmutter et al., 1999, Experiment 3), all of them using a PP design. This inconclusive pattern could point to two competing factors driving the effect into opposite directions. The same competition should happen in plural-verb *ungrammatical* conditions, too, but here the overall pattern is clearly facilitatory. This could, however, be explained by 1) the fact that, in ungrammatical conditions, there are more studies testing other designs than the PP construction, and 2) that effects in ungrammatical conditions are usually stronger and could thus overrule a spilled-over effect of distractor number.

Apart from grammatical role, discourse topicality, and linear order, which we recorded in the review, there might be other contributing factors that we have not considered here: Factors like thematic role (Arnold, 2001), contrastive focus (Cowles, Walenski, & Kluender, 2007), first mention (Gernsbacher & Hargreaves, 1988), and animacy (Fukumura & van Gompel, 2011) are known to affect discourse saliency and might thus influence distractor prominence.

3.5.2 Cue Confusion

While the influence of distractor prominence on effect size is an intuitively comprehensible assumption and has much independent empirical support, the concept of cue confusion we introduced is a novel theoretical construct that, to our knowledge, has never been investigated in sentence comprehension research. Cue confusion is motivated by independent principles such as frequency learning and associative cues. There are currently only a few studies that independently motivate the proposal. The data available in support for cue confusion are from Mandarin reflexives and Hindi reciprocals. The inhibitory effects found could thus be related only to certain language-specific characteristics and not differences in feature-co-occurrence. More research on this topic is needed in Hindi and Mandarin and also on English reciprocals.

A possible way to test the cue confusion hypothesis for English in a controlled experiment would be to directly compare reflexives and reciprocals, manipulating the number cue in both. An example design we have also suggested in Jäger, Engelmann, and Vasishth (2015) is shown in (18).

(18) a. *Reflexive; distractor-match*
The *nurse* who cared for the *children* had pricked *themselves* …

b. *Reflexive; distractor-mismatch*
The *nurse* who cared for the *child* had pricked *themselves* …

c. *Reciprocal; distractor-match*
The *nurse* who cared for the *children* had pricked *each other* …

d. *Reciprocal; distractor-mismatch*
The *nurse* who cared for the *child* had pricked *each other* …

Under the cue confusion hypothesis, a reduced facilitatory or even inhibitory effect is predicted for the reciprocal *each other* compared to the reflexive *themselves.* In order to derive a finer-grained metric that predicts differences in cue confusion levels between different dependency environments, co-occurrence frequencies could be computed from a corpus in which sufficient dependency information is available.

3.5.3 Cue Strength

Another factor that can affect the result of an experiment is the choice of the cue that is manipulated. Several authors have considered that syntactic cues might be weighted more strongly than semantic cues in selecting a target for retrieval (e.g., Nicol & Swinney, 1989; Sturt, 2003; Van Dyke, 2007; Van Dyke & McElree, 2011). However, besides the syntax/semantics distinction there are more properties one can use to categorize cues or features. For example, while number is a grammatical feature which, in English, is overtly marked, animacy and gender are lexical and mostly not orthographically overt. Person is supposedly a strong manipulation since it affects the speaker/listener perspective. Some features rather express a semantic appropriateness for a verb instead of strict requirements: For example, in the experiment design of Van Dyke and McElree (2006), the manipulation was whether the distractors had a specific property, e.g., being *fixable*. In addition to markedness asymmetries within features such as plural vs. singular, a hierarchy between features (person > number > gender) has been proposed, e.g., by Carminati (2005). Finally, binary features such as c-command, subject, or animacy could be stronger than features with more possible values such as gender or case.

It has also been proposed that the weighting of cue strength differs between dependency types. Effects are harder to detect in reflexive-antecedent dependencies than in number agreement. In particular, (Dillon et al., 2013; Kush & Phillips, 2014; Nicol & Swinney, 1989; Parker & Phillips, 2014; Phillips et al., 2011; Sturt, 2003; Xiang et al., 2009) have claimed that structural cues have priority over semantic cues in the processing of reflexives but not in the processing of agreement in order to explain why effects are harder to detect in reflexive-antecedent dependencies than in than in number agreement.

The experiments and simulations by Parker and Phillips (2014) that we discussed above support the idea of priority for structural cues specifically for reflexives. However, simulations with our extended model revealed that differently weighted cues are not a necessary assumption in this case. It is nevertheless possible that cue strength difference between cues and even between dependencies. Further research is necessary on this matter, since the data available to date is not conclusive.

Note also that the lower significance rate in reflexives compared to number agreement could in several cases be accounted for by an alternative explanation instead of syntactic priority: In studies on reflexives, the cue manipulation is in many cases not as strong as it is in agreement studies: The majority of experiments in reflexive processing violates the *stereotypical* gender of the target in target-mismatch conditions, e.g., using *surgeon* with *herself*. However, this can only account for those experiments that manipulated the gender cue. A counter-example is Dillon et al. (2013), who compared interference effects in agreement and reflexives manipulating the *number* cue in both constructions and found no effect in reflexives while observing a clear facilitation in agreement.

In any case, it is questionable whether one can directly compare the effects of number agreement and reflexives-antecedent dependencies as long as it is not clear what is the cause for the unique pattern of effects in target-match conditions of number agreement.

3.5.4 Quality of the Data

Due to the general sparsity of the data, a full statistical evaluation of our proposals is not yet possible. Furthermore, many (if not all) studies discussed have remarkably low statistical power; in these studies, power tends to lie in the range 0.05-0.20. As Gelman and Carlin (2014) point out, even ignoring the usual problems of Type I and II errors, if statistical power is as low as 0.05, the sign of the effect could be wrong as much as 30% of the time (Type S errors) and the magnitude of the effect can be radically exaggerated, e.g., by a factor of 7 (Type M errors). The power problem is a very general one in psycholinguistic research, and especially sentence processing, which investigates phenomena with very small effect sizes. As a consequence of these statistical issues, the researcher can respond in several different ways to deviations of experimental results from theoretical predictions in the published literature. In our specific case, the literature shows facilitatory effects in target-match conditions and inhibitory effects in target-mismatch conditions, and the ACT-R-derived standard assumptions about cue-based retrieval cannot explain these effects. One response from the researcher can be to treat all these results as noise and just stick to the standard theory. Another response can be to assume additional mechanisms that only apply to these specific cases, possibly overfitting to the data. A third option is to take all the unexplained data seriously and consider the classical account as falsified and wrong in its entirety and just stop there. In the present work, we try to take both the theory and the data seriously. We suggest that the apparently inconsistent results possibly constitute systematic variation that could be explained by characteristic differences between individual studies. We attempt to develop a generally applicable theory that could plausibly account for many of the effects. The claims in the work presented here are falsifiable, and rigorous tests of the extended theory should be aggressively pursued in future work, although of course such an investigation would be more informative if conducted with appropriately powered studies.

3.5.5 Generality of the Extended Model

The extended model predicts effects of distractor prominence and cue confusion by the implementation of two independently motivated principles: The prominence correction and the concept of associative retrieval cues. The prominence correction implements the intuitive prediction that the relative activation of a competitor with respect to the retrieval target should influence the strength of similarity-based interference. In the cognitive modeling framework ACT-R, which the Lewis and Vasishth (2005) model is based on, the implementation of SBI (the ACT-R fan effect) does not take into account the activation difference between items. We believe that this is not intentional, but that predictions with respect to this feature have simply not been part of computational research using ACT-R so far. We also think that the principle of associative retrieval cues is not specific to language but a general feature of cue-based retrieval in human cognitive processing. We therefore propose it to be integrated in the widely-used framework

ACT-R. This will help to promote further research on cue confusion in general cognition.

3.5.6 Conclusion

Based on a large-scale literature review and computational modeling of retrieval interference in dependency resolution, we arrive the following conclusions. The mechanisms underlying the processing of number agreement, in particular in target-match conditions, differ from those underlying the processing of other types of dependencies. In contrast, non-number subject-verb dependencies and reflexives-antecedent dependencies can be explained by the same mechanisms. We have identified distractor prominence and cue confusion as determinants that predict variability across results reported in the literature on retrieval interference. In many cases, the failure to find effects and even the inversion of the direction of effects in target-match conditions can be explained by distractor prominence. Inhibitory effects in target-match conditions we propose to be explained by the confusion of retrieval cues, resulting from cue/feature co-occurrence frequencies.

We have not found results that clearly support the proposal that syntactic cues serve as a filter or are prioritized specifically in the processing of reflexives. We, nevertheless, acknowledge that cues might have variable strength across different cue types and also across dependencies, as this theoretically follows from our proposal of associative cues.

The extended model of cue-based retrieval, for the first time, provides quantitative predictions with respect to systematic variabilities in experimental design across studies. The presented model is therefore an important step forward in helping us interpret results in the context of previous findings and for formulating computationally informed predictions for future experiments.

The two principles of prominence correction and associative cues that constitute our extended model are compatible with the general theory of cue-based retrieval as the essential mechanism underlying dependency resolution in sentence processing. Both principles are independently motivated and should be considered as domain-general mechanisms and as extensions to current cognitive frameworks such as ACT-R.

We strongly encourage further research in order to support or falsify our claims about distractor prominence and cue confusion in language processing and in general cognition. Other researchers are invited to use the extended model presented here to do further simulations. As a tool for this purpose, we provide an online application of the model at `https://engelmann.shinyapps.io/inter-act`.

Chapter 4

A Basic Interface between Parsing and Eye Movement Control

The contents of this chapter are published in:

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, *5*, 452–474. doi: 10.1111/tops.12026

4.1 Introduction

In order to understand how post-lexical processing and eye movements interact, it is necessary to combine computational models of these two areas and investigate the link between high-level language processes and oculomotor control. In a recent approach, Reichle et al. (2009) introduced a post-lexical integration stage into E-Z Reader 10, that interacts with eye movement control through regressions. Whenever the integration stage takes too long, a regression is triggered in order to buy time for the integration process to finish. Although Reichle and colleagues did not integrate a computational account of post-lexical processing, they showed a suitable way toward studying the link between parsing and eye movements.

In the work presented here, the cognitive architecture ACT-R (Anderson et al., 2004) is used to combine an eye movement control model with a parser in a similar way as Reichle et al. (2009) did. However, we incorporate two well-tested computational accounts of parsing difficulty that capture memory retrieval and structural prediction, respectively: (1) The cue-based retrieval parsing account of Lewis and Vasishth (2005) builds on independently motivated assumptions about memory access and has been implemented as a fully specified parser in ACT-R; (2) Surprisal

(Hale, 2001; Levy, 2008) defines difficulty in terms of disconfirmed structural predictions. The combination of both metrics in one model is motivated by empirical evidence and statistical modeling: Experimental results suggest a complementary relation between expectation-based and working-memory-based accounts (Demberg & Keller, 2008; Konieczny, 2000; Staub, 2010a; Vasishth & Drenhaus, 2011), and corpus studies show that surprisal and retrieval are independent predictors of processing difficulty (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Boston et al., 2011; Patil, Vasishth, & Kliegl, 2009; Vasishth & Lewis, 2006). The use of ACT-R has several advantages. First, ACT-R implements cognitive principles that are valid in distinct domains and enables the development of models for various tasks. Second, it integrates all levels of cognition from visual and motor processes that interact with a virtual outside world to rule-based reasoning. Third, ACT-R is a model of real-time processing, which makes its predictions directly comparable to eye-tracking data in milliseconds. As eye movement model we use the ACT-R-integrated EMMA ("eye movements and movement of attention", Salvucci, 2001), which is in principle a simplified and domain-independent version of E-Z Reader.

The goal of this chapter is to demonstrate the feasibility of integrating a computational account of post-lexical difficulty with an eye movement control model. For that purpose, we avail ourselves of a framework which is simplifying in some respects but exhibits enough flexibility for further development and extension. In order to provide a general assessment which is comparable to earlier studies (Reichle et al., 1998, 2009; Salvucci, 2001), we perform a qualitative examination of the framework on a suitable eye-tracking corpus. Although E-Z Reader and EMMA were evaluated on the Schilling Corpus (Schilling, Rayner, & Chumbley, 1998), we used the German Potsdam Sentence Corpus (Kliegl et al., 2004) because the two metrics of parsing difficulty surprisal and retrieval are readily available for the latter.

Sections 4.2 and 4.3 will introduce E-Z Reader and EMMA, respectively. In Section 4.4, we present a replication of Salvucci (2001) on the English Schilling Corpus, which is necessary because ACT-R has developed further since Salvucci's evaluation of EMMA in 2001, and EMMA itself has been re-implemented. The successful replication provides the basis for an extension of the model with parsing theory, which will be described in Section 4.5. Finally, Section 4.6 presents six simulations on the German Potsdam Sentence Corpus that assess a range of model configurations that integrate EMMA with surprisal and memory retrieval.

4.2 E-Z Reader

Besides SWIFT (Engbert et al., 2005), the E-Z Reader model of eye movement control in reading (Reichle et al., 1998; Reichle, Pollatsek, & Rayner, 2006, 2012; Reichle et al., 2009) is the most successful model of the detailed *when* and *where* of the eyes during reading. It predicts well-established observations such as effects of lexical frequency and predictability on preview and word skipping or the influence of saccade length on landing site distributions and refixations. As this model serves as inspiration for EMMA and for the eye-parser interface presented here, it will be described briefly below.

In E-Z Reader (EZR), attention is allocated serially from word to word and is partly decoupled from eye movement control. Word identification consists of two stages: A *familiarity check* initiates the programming of a saccade to the next word and the completion of *lexical access* triggers the attention-shift to the next word. The duration of the first stage, the familiarity check ($L2$), is influenced by the word's corpus frequency (e.g., Francis & Kucera, 1982), length, and cloze predictability (Taylor, 1953) in the way that more common and more predictable words are recognized faster. The second stage, lexical access ($L2$), is defined as a fraction of $L1$ ($0.25L1$ in EZR 10) and represents the activation of the word meaning. Saccade programming also consists of two stages: The first stage ($M1$) is a *labile* stage which can be canceled when a new saccade program is initiated. This happens when the next word is easy to identify so that the familiarity check $L1$ completes before $M1$ and results in *skipping* of the next word. Its duration is about 125 ms. The second stage ($M2$) is non-labile; once initiated, it cannot be interrupted. Its duration is about 25 ms. The actual saccade execution is also fixed to 25 ms. Due to systematic and random error, saccades overshoot and undershoot their targets, resulting in corrective refixations.

EZR 10 contains a post-lexical integration stage (I), which serves as a placeholder to "reflect all of the post-lexical processing necessary to integrate word n into the higher-level representations that readers construct on-line; for example, linking word n into a syntactic structure, generating a context-appropriate semantic representation, and incorporating its meaning into a discourse model" (Reichle et al., 2009, p. 7). It starts after completion of word identification and, in most cases, does not influence fixation durations or saccade planning. It interferes with eye movement control only in the case of an error. As Reichle (2015) put it, "lexical access is like 'stepping on the gas pedal' to move attention and the eyes forward, whereas post-lexical processing is like 'not stepping on the brakes', allowing attention and the eyes to continue their progression" (p. 279).

Figure 4.1 from Reichle et al. (2009) shows three possible processing sequences in EZR 10. Panel (A) shows the usual sequence when word integration (I) with mean length of 25 ms proceeds without errors. In this case, post-lexical processing has no influence on saccade timing. However, word integration can be delayed due to difficulty, which is exemplified in panel (B). In this case, at completion of lexical access of the next word, a regression is programmed toward the critical word, where the difficulty appeared (or, with some probability, an earlier region), and post-lexical processing is restarted. This is called "slow integration failure". Another type of integration error occurs when a word violates structural expectations and disrupts the integration process. This violation would be detected early and is therefore called "rapid integration failure". As shown in panel (C), this interrupts the programming of the forward saccade and directs attention back to the critical word. In contrast to slow integration failure, the case of a rapid disruption more

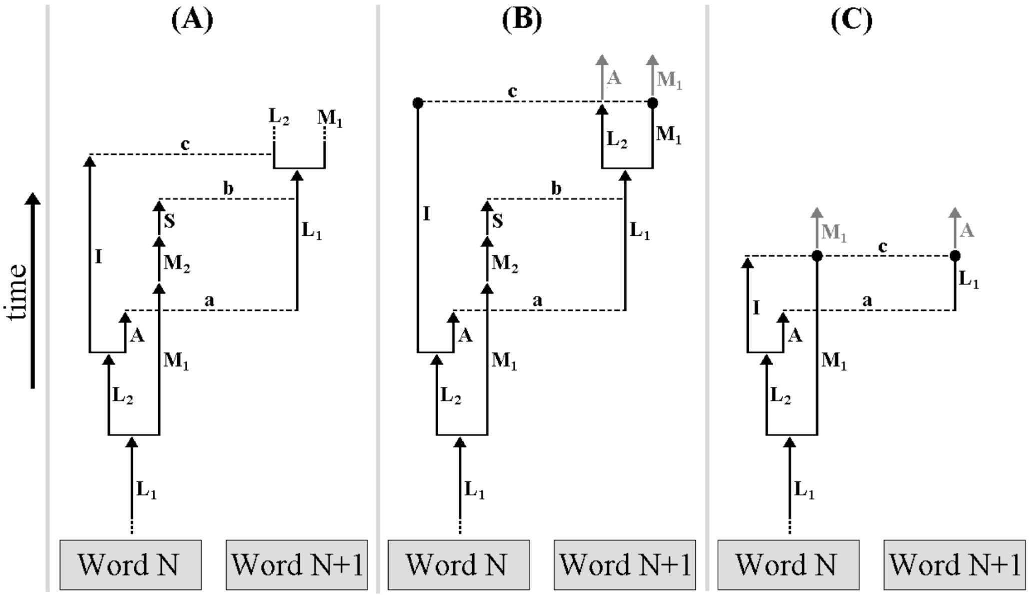

Figure 4.1: Figure 2 of Reichle et al. (2009). Three sequences of events in E-Z Reader 10.

frequently results in inflated fixations instead of a regression, because the forward movement is halted before the eyes have moved on to the next word.

As a demonstration of the applicability of the post-lexical extension, Reichle and colleagues simulated effects of clause wrap-up, semantic violations, and garden-paths by assuming certain probabilities of slow or rapid integration failure in each of the three cases. Thus far, parsing in EZR is only represented as a failure probability. The next step is now to fill this placeholder with an explicit processing theory that predicts parsing duration and failure.

4.3 The EMMA Model (Salvucci, 2001)

EMMA's basic assumptions were inspired mainly by E-Z Reader. The main characteristics of the model are a dynamic calculation of word encoding time and a distinction between overt eye movements and covert shifts of attention. Attention is allocated serially and proceeds usually ahead of the eye movement. This enables the model to produce skipping and refixations. The programming of saccades consists of a labile stage, i.e., a stage that can be canceled by upcoming attention shifts, and a non-labile state, after which the saccade preparation has passed a point of no return leading to an eye movement inevitably. Below we describe the version of EMMA that we used for our simulations in the environment of ACT-R 6.0.

The core function of EMMA calculates the encoding time of an object based on its frequency of occurrence and its eccentricity from the current viewing location. The resulting duration represents attention shift and word identification in one step. The encoding time T_{enc} is calculated in the following way:

$$T_{enc} = K(-\log f_i)e^{k\epsilon_i} \tag{4.1}$$

where K (visual encoding factor) and k (encoding exponent) are scaling constants, ϵ_i is the eccentricity of the object (i) to be encoded, and f_i is the object's corpus frequency normalized to a range between 0 and 1 (word occurrence per one million words divided by one million). The saccade preparation time T_{prep} has been estimated in Salvucci's simulations to 135 ms.[1] The non-cancelable stage T_{exec} consists of 50 ms for saccade programming, 20 ms for saccade execution and additional 2 ms per degree of visual angle of the saccade length. The model introduces variability to T_{enc}, T_{prep}, and T_{exec} by randomly drawing from a uniform distribution[2] with a standard deviation of one third of the actual value. Also, landing point variability of a saccade is defined by a Gaussian distribution with a standard deviation of 0.1 times the intended saccade distance. For empirical motivations for the choice of distributions, see Salvucci (2001).

Salvucci presented three evaluations of his EMMA/ACT-R model on empirical data from equation-solving, visual search, and reading. In the case of reading, which is the application of interest here, EMMA was interfaced with a simple ACT-R model that worked in the following way: Each cycle begins with the initiation of an attention shift to the nearest object to the right. EMMA then initiates the encoding of the target object using the provided frequency values and, at the same time, starts the preparation of the corresponding eye movement. Once the visual encoding has finished, the model performs a lexical retrieval of the input word and starts the next cycle by shifting attention to the next word. The lexical retrieval had a fixed duration and, thus, did not contribute to the predictions in a relevant way. Salvucci tested EMMA on the 48 sentences of the Schilling Corpus (Schilling et al., 1998) and showed that the model reproduced well-known empirical effects of word-frequency on a range of eye-tracking measures.

4.4 Replication of Salvucci (2001)

4.4.1 Method

Data The Schilling Corpus (SC) contains fixation data of 48 American English sentences with 8-14 words each, read by 48 students. For evaluating the model performance, Salvucci (2001) used data compiled by (Reichle et al., 1998). They had calculated the means of six eye-tracking measures for five logarithmic frequency classes (see Table 4.1). The frequency values available in

[1] In ACT-R 6.0, the planning time for motor processes amounts to 0, 50, 100, or 150 ms depending on feature-based similarity with the previous movement. However, for our simulations we used Salvucci's original definition of a fixed preparation time.

[2] A uniform distribution is the ACT-R 6.0 default for random time generation. In Salvucci's original model a Gamma distribution was used.

Table 4.1: Frequency classes used in the analyses of the Schilling Corpus (SC) and Potsdam Sentence Corpus (PSC)

		SC		PSC	
Class	Freq. in 1M	Words	Mean	Words	Mean
1	1–10	77	3	186	3
2	11–100	87	50	173	41
3	101–1000	71	333	200	335
4	1001–10,000	92	5067	207	5020
5	>10,000	112	41976	84	2399

the SC were obtained from Francis and Kucera (1982). In order to avoid confounds, the first and the last word of each corpus sentence was removed. Since the model did not produce regressions, trials that contained inter-word regressions (64%) were excluded from the analysis.

Model Our ACT-R model consisted of four productions: `find-next-word` (search for the nearest object to the right), `attend-word` (initiate an attention shift and encoding by EMMA), `integrate-word` (start memory retrieval), and `stop-reading` (when the sentence is finished). The `integrate-word` rule did not do anything in this model apart from adding 50 ms to the processing time. It was used in later simulations, however, to initiate the parsing process. All simulations presented here were carried out in ACT-R 6.0. We used EMMA version 4.0a1 (with some minor adjustments by us) as it has been re-implemented by Mike Byrne and Dan Bothell in order to be fully integrated in ACT-R 6.0. All parameters except for those shown in Table 4.2 were kept at their default values. This is particularly important for the *default action time*, which is the firing duration assigned to each ACT-R production rule. Salvucci (2001) set it to 10 ms, but in ACT-R 6.0 it defaults to 50 ms.

Analysis One simulation consisted of 10 complete model runs through the 48 sentences of the Schilling Corpus. Fixations times were recorded for each word. The analysis was carried out in the R statistics software (R Core Team, 2012). Following the analysis of (Reichle et al., 1998) and (Salvucci, 2001), we excluded first and last words from the sentences and all trials that contained inter-word regressions. Then we divided the corpus words into five logarithmic frequency classes (see Table 4.1) and calculated the means for each class for six fixation measures: gaze duration (the time spent on a word during first pass, including immediate refixations), first fixation duration (FFD, duration of the first fixation on a word during first pass), single fixation duration (SFD, fixation duration on a word if it is fixated only once during first pass), the skipping rate of a word (skip), the probability of fixating a word exactly once (onefix), and the probability of fixating a word more than once (refix). This analysis was done with both the experimental data and the model output. We quantified the goodness of fit between the model predictions and the data using the *Pearson product-moment correlation coefficient R* and the

Table 4.2: Fit and parameter estimates for all simulations

			Parameters				Fit			
			K	T_{prep}	F	P	R_{early}	R_{late}	$RMSD$	$\%reg$
no regr.		Salvucci (2001)	0.006	0.135			0.97		0.362	0
	1	SC replication	0.002	0.135			0.96		0.303	0
	2a	PSC	0.003	0.120			0.86		0.326	0
PSC all	2b	EMMA	0.003	0.120			0.91	0.38	0.638	0
	3	+s$_1$	0.002	0.115		0.0030	0.93	0.39	0.645	0
	4	+r	0.002	0.110	0.2		0.90	0.86	0.201	18
	5	+s$_2$	0.003	0.115		0.0200	0.92	0.87	0.229	15
	6	+rs$_1$	0.003	0.115	0.2	0.0005	0.92	0.88	0.257	12
	7	+rs$_2$	0.003	0.115	0.1	0.0150	0.90	0.91	0.206	23

Notes: K = EMMA encoding factor, T_{prep} = EMMA saccade preparation time, F = ACT-R retrieval latency factor, P = scaling factor for surprisal. The fit was calculated for means of 5 frequency classes for each eye-tracking measure. R_{early} = correlation coefficient between observed and predicted values for early measures (gaze, FFD, SFD, skip, onefix, and refix). R_{late} = correlation coefficient for late measures (RPD, TFT, RRT, FPREG, and reread). The last two columns show the total normalized root-mean-square deviation and the percentage of simulated trials that contained regressions.

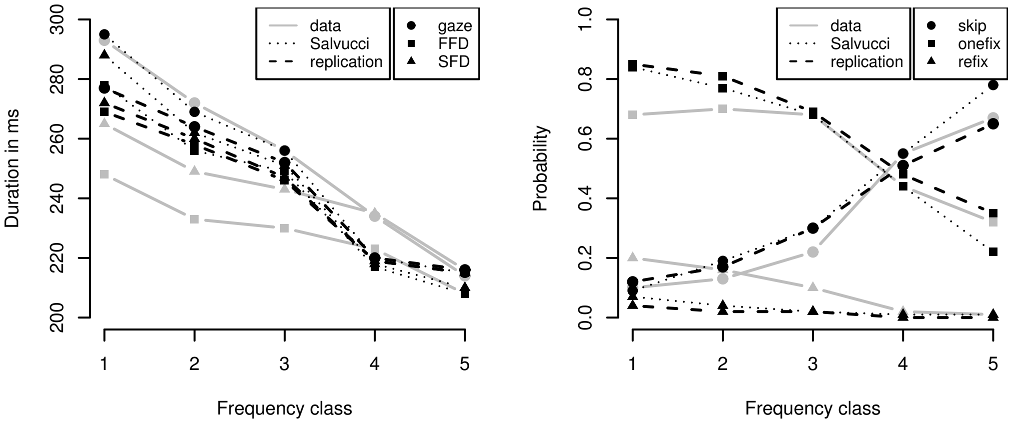

Figure 4.2: Replication of Salvucci (2001) on the Schilling Corpus. Effects of word frequency on gaze, first, and single fixation duration, and on the rate of skipping a word, fixating it once and fixating it more than once. Grey solid lines represent experimental data, black dotted lines show Salvucci's simulation results, and black dashed lines show the replication results. Lexical frequency is divided into classes 1 (lowest) to 5 (highest).

root-mean-square deviation ($RMSD$). $RMSDs$ were normalized by the standard deviation of the observed data in the same way as it was done in (Reichle et al., 1998) and Salvucci (2001). A precise definition is given below.

Root-mean-square deviation The root-mean-square deviation ($RMSD$) is used to estimate the relative goodness of fit between predicted and observed data. Reichle et al. (1998) and Salvucci (2001) normalized the $RMSD$ to be comparable between different scales (milliseconds and probabilities) by dividing the difference between observed and predicted values by the standard deviation of the observed values. In their Appendix, Reichle et al. state that this normalization was done after squaring the difference. However, the actual $RMSD$ values in Reichle et al. (1998) and Salvucci (2001) were obtained by first dividing the difference by the standard deviation and then squaring it.[3] For the reason of comparability we also used the latter definition. For each model, we calculated the $RMSD$ for the frequency statistic over all fixation measures and frequency classes as defined below:

$$RMSD = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{data_k - model_k}{SD_k}\right)^2} \tag{4.2}$$

where $data_k$, $model_k$, and SD_k range over all fixation measures and frequency classes.

The parameter optimization procedure was carried out by first identifying a number of parameter configurations with R values near the maximum and then, among these, choosing the one with the smallest $RMSD$. In this way, the optimization represented a priority for the quality of effects while also taking quantity into account.

4.4.2 Results

We re-estimated the encoding factor K and the saccade preparation time T_{prep} in order to compensate for the changes in the ACT-R environment. See Table 4.2 for a summary of the simulation results including estimated parameter values. The parameter fitting resulted in a decrease of K, which should mainly be due to the increased default action time of 50 ms in ACT-R 6.0. Fig. 4.2 shows the predictions of the model (dashed lines) for six fixation measures as a function of frequency class. Besides the corpus data (grey solid lines), we also plotted the results of the original study (dotted lines) as reported in Salvucci (2001) for comparison. The main trends in the data are that high-frequency words are read faster and skipped more often than low-frequency words. These trends and the overall pattern of the data were reproduced by the model with a close fit to the original predictions. The mean correlation R with the data was 0.96 and the mean $RMSD$ was 0.303.

[3] We concluded this by a recalculation of their values and personal communication with Dario Salvucci.

4.4.3 Discussion

The EMMA/ACT-R model, as re-implemented in ACT-R 6.0, reproduces frequency effects on fixation durations and probabilities in the Schilling Corpus with a performance comparable to that of the original simulation of Salvucci (2001). Despite the different environment, a small adjustment to the encoding time was sufficient to replicate the results. The successful re-evaluation of EMMA in its current version is essential for the next steps that will extend the model with accounts of parsing theory.

4.5 Interface I: Time Out

In order to augment the EMMA/ACT-R model with post-lexical processing, we take a similar approach as Reichle et al. (2009). The integration stage of E-Z Reader 10 operates in parallel to eye movement control but can interrupt the reading process for two reasons: Either integration of a word w_n just fails ("rapid integration failure") or the integration process takes too long ("slow integration failure"), which means that integration of word w_n does not finish before identification of word w_{n+1} is completed. In either case the eyes are directed back to word w_n or w_{n-1} with a certain probability.

Our goal is to evaluate a model which works in a similar way but uses a computational theory of sentence comprehension to generate its predictions. Our implementation is similar to slow integration failure, which Reichle et al. (2009) describe in the following way:

> ... word n+1 is identified (i.e., L2 completes) before word n has been integrated, which ... halts both the post-lexical processing (I) of word n and the forward movement of the eyes (e.g., in this example, M1 is canceled) so that both attention and the eyes can be directed back to the source of processing difficulty. (p. 10)

Similarly, in our model, when identification of word w_{n+1} finishes before the complete integration of word w_n, a regression back toward the previous word is initiated. However, different from EZR 10, post-lexical processing is not halted but continues during the regression. Thus, our implementation is comparable to the mechanism that Mitchell et al. (2008) have described in their "Time Out hypothesis", which assigns to these short regressions the function to provide additional time for the sentence processor to catch up with a backlog before taking up new input. Once word integration is complete, the model continues with normal reading.

Figure 4.3 shows how, in our model, the ACT-R-based parsing architecture of Lewis and Vasishth (2005) interfaces with the EMMA eye movement control model. The upper panel shows the uninterrupted reading process. The only ACT-R rule that interacts with eye movement control in a top-down way is the shift of attention. As soon as an attention shift is requested (ATTENTION) the eye movement module starts the word recognition process (ENCODING) and at the same

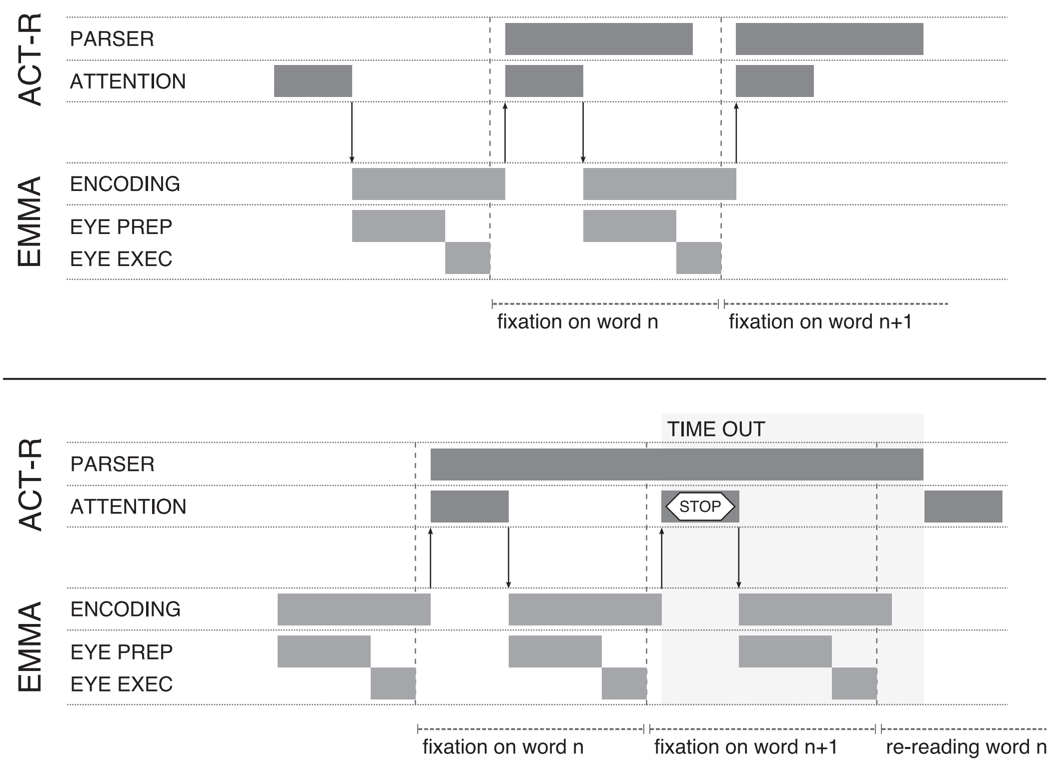

Figure 4.3: Illustration of the Time Out function. The upper panel shows an uninterrupted sequence of reading word w_n and word w_{n+1}. The lower panel shows the situation where integration of word w_n is delayed and causes a regression.

time programs a saccade to the same word. The preparation stage of the saccade programming (EYE PREP) can be canceled by an upcoming attention shift, which leads to a skipping of the targeted word. Once the beginning of the execution stage (EYE EXEC) has passed, an eye movement will be carried out inevitably. The completion of the attention shift, which includes the recognition of the word, is the signal for the parsing module (PARSER) to begin the integration into the syntactic structure. This includes the creation of new syntactic nodes, the retrieval of previously created structural chunks from memory, and finally the grammatical combination of both. While the parser is carrying out these steps, attention is shifted to the next word and a new saccade is programmed. The time needed to retrieve an item from memory varies as a function of decay over time and similarity-based interference. Consequently, dependent on the syntactic configuration of the sentence, it is possible that the structural integration of a word is still in process while the recognition of the next word has already completed. This scenario is shown in the lower panel. In this situation, the next word naturally cannot be integrated yet. Instead a time-out rule fires, which initiates an attention shift to the left of the current word in order to buy time for the integration process to finish.

Since, in ACT-R, only one retrieval request can be handled at a time, it follows naturally that retrieval of word w_n has to be completed before the integration of word w_{n+1} can start. Once initiated, the retrieval process operates in parallel to cognition and eye movement planning. As long as the difficulty is low and retrieval completes fast, the reading process is uninterrupted.

The concept of interrupting the "normal" reading process by time-outs should not be misunderstood in the way that making regressions is not normal. We assume that these interruptions by the parser belong to normal reading as they happen quite regularly and are not under conscious cognitive control. A quite different case are active reanalysis mechanisms where the reader is aware of an inconsistency (syntactic or semantic) and has to make long-range regressions. However, although the presented framework can be used to study this kind of behavior, we restrict our study to the simplest case for now.

In addition to observable regressions, the mechanism described above predicts increased fixation durations as a by-product: When the delayed integration process finishes during the labile stage of the eye movement program before the regression is executed, the current program is canceled and a new forward saccade is planned. In the behavioral record, this scenario appears just as inflated reading time.

For simulating post-lexical processing, we use two complementary explanations of parsing difficulty: Cue-based retrieval (Lewis & Vasishth, 2005) and syntactic surprisal (Hale, 2001; Levy, 2008).

The sentence processing model of Lewis and Vasishth is a fully-specified parser the actions of which can be transparently measured in milliseconds. It relies on domain-independent memory principles, and it is well-tested by a number of applications. This kind of model is exactly what is needed in order to investigate the interaction between parsing and eye movements in detail. We connect this parsing model to EMMA through the Time Out mechanism.

Surprisal (Hale, 2001; Levy, 2008) formalizes the idea that unexpected structures cause processing difficulty (Konieczny, 2000). Hale defined the surprisal of a word as a function of the probability mass of all derivational options that have to be disconfirmed at that point in the sentence. The surprisal of a word w_i is the negative logarithm of the transition probability from word w_{i-1} to w_i. The lower the probability of a word given its preceding context, the higher its surprisal. While Hale assumed a complete knowledge of the grammar to define the surprisal value, there are also different accounts of calculating surprisal, e.g., using a neural network (Frank, 2009) or using a rationally bounded parallel dependency parser (Boston et al., 2011).

Although the difficulty associated with surprisal stems from building low-probability structures, it is not clear that the cause of the difficulty must be located in post-lexical processing. Given the conceptual distinctness of surprisal and retrieval together with experimental evidence locating expectation effects earlier than memory effects (Staub, 2010b; Vasishth & Drenhaus, 2011), we hypothesize that the source of these two types of difficulty may lie at different points in the processing time course. Theoretically, it is legitimate to assume that the contextually pre-activated high-probability structures (or parsing steps) would also pre-activate lexical items belonging to the according categories. In that case, at every point in the sentence the activation of specific lexical items receives a boost by its structural context. This would directly affect the speed of the word identification process. That means, although the source of surprisal difficulty is undoubtedly a "high-level" post-lexical process, the actual realization of that difficulty could happen "low-level" at the stage of word identification.

The following simulations test both assumptions, surprisal affecting the high-level and affecting the low-level. The high-level variant is implemented by additively modulating the duration of the integration stage by a scaled surprisal value. For simulating surprisal affecting the low-level we include the surprisal values in EMMA's core equation of word encoding time. The resulting equation for T_{enc} will be shown in the next section.

4.6 Evaluation on the Potsdam Sentence Corpus

In this section, we present six simulations that were carried out on the Potsdam Sentence Corpus (PSC, Kliegl et al., 2004). The PSC was used because Boston et al. (2008) and Boston et al. (2011) provide retrieval and surprisal values for all corpus words. Simulation 2 evaluated EMMA on the PSC in order to compare the results with the model performance on the Schilling corpus. Besides assessing how well the model can be generalized to another corpus in a different language, this study pursued the goal to establish the basis for augmenting the EMMA/ACT-R model with post-lexical processing. The other five simulations tested EMMA in different configurations that include and combine retrieval (r), low-level surprisal (s_1), and high-level surprisal (s_2): EMMA+s_1, EMMA+r, EMMA+s_2, EMMA+rs_1, and EMMA+rs_2 (see Table 4.2 for an overview).

4.6.1 Method

4.6.1.1 Potsdam Sentence Corpus

The Potsdam Sentence Corpus contains eye-tracking data from 144 simple German sentences (1138 words) with 5 to 11 words per sentence, read by 229 readers. The corpus contains values of printed word frequency obtained from the CELEX database, a corpus of about 5.4 million words (Baayen, Piepenbrock, & van Rijn, 1993). Kliegl et al. (2004) report effects of frequency on reading times and probabilities using the same logarithmic frequency classes that were used in Salvucci (2001) (see Table 4.1). The trends are comparable to those in the Schilling Corpus: Higher frequency correlates with shorter reading times and higher skipping rates, although the trend is not as strong in first and single fixation durations.

We integrated retrieval and surprisal information in the corpus data that provided the input for the EMMA/ACT-R model.

4.6.1.2 Retrieval

There are handcrafted ACT-R parsing rules available for a number of psycholinguistically interesting sentence constructions; however, not enough to cover the whole PSC. For this corpus-based benchmarking evaluation carried out here, we therefore used pre-calculated values from Boston et al. (2011). These retrieval values were calculated using a parallel dependency parser and approximately represent the duration a retrieve-and-attach cycle would require in the ACT-R parser. Each step of the dependency parser (`SHIFT`, `REDUCE`, `LEFT`, `RIGHT`) was assigned a duration of 50 ms — the *default action time* in ACT-R that it takes one production to fire. The duration of retrieving an item from memory was calculated using ACT-R equations, including a simplified version of similarity-based interference. The parser was assessed at different levels of parallelism, i.e., the number of alternative derivations to be pursued at the same time. The retrieval values obtained at the highest level of parallelism (100 parallel analyses) were the most significant predictors in Boston et al. (2011). These values (M = 357.8 ms, SD = 122.16 ms) were used in our model to imitate the parsing process. The values were scaled with the ACT-R-internal *retrieval latency factor* F.

4.6.1.3 Surprisal

For the present purposes, we used surprisal values (M = 2.9 bits, SD = 2.06 bits) from Boston et al. (2008), which were generated with a modified version of the probabilistic context-free phrase-structure parser[4] from Levy (2008).

4.6.1.4 Model

For the following simulations, the model used in the replication of Salvucci (2001) was modified in the way described in the previous section. After encoding word w_n, the `integrate-word` rule starts the parsing actions and attention is shifted to the next word to the right. For the current study, the parsing duration was imitated by a timer set to the corresponding retrieval value from Boston et al. scaled by F. As long as the timer is running, no other word can be integrated.

In order to establish a link between cognition and eye movement control, two ACT-R production rules were added to the model: `time-out` and `exit-time-out`. Their function is as follows: When integration of word w_n is still in progress while the encoding of word w_{n+1} has already

[4] The parser is publicly available at http://nlp.stanford.edu/∼rog/prefixparser.tgz

completed, `time-out` initiates an attention shift to the word to the left of the currently fixated one (Time Out regression). Once integration of word w_n has finished, the `exit-time-out` rule returns the model into the state of normal reading. For reasons of simplicity, no special assumptions are made about the reading process just after a Time Out regression, except for the fact that word w_n will not need to be integrated again. However, word w_n and w_{n+1} will go through the identification process again after leaving Time Out mode because word encoding is part of every attention shift carried out by EMMA. A more realistic model would probably not fully re-encode a word already identified.

Note that a Time Out regression can be initiated from word w_n or w_{n+1} depending on how fast the encoding process of word w_{n+1} is in relation to the saccade execution to that word. The regression always targets the word to the left of the current fixation. This means, the regression target can either be word w_n or w_{n-1}. However, the preparation of a regression can be canceled before its execution in the case when the integration process completes before the non-cancelable state of motor preparation. In this case, the time out would show itself in the form of a refixation on w_n or w_{n+1}. In case this refixation is also canceled because encoding was fast, a saccade to the next word is planned and the time out only causes an increased fixation duration.

Finally, we included the two versions of surprisal described above. We equipped ACT-R with a surprisal scaling constant P. For simulating surprisal at the high level, the values scaled by P were added to the duration of the integration stage in milliseconds. In order to modulate the low-level word encoding process directly by surprisal, we added surprisal in EMMA's word encoding time equation as shown in Equation 4.3:

$$T_{enc} = (K[-\log f] + Ps)e^{k\epsilon} \tag{4.3}$$

where s is the surprisal value of the corresponding word, and P is the surprisal scaling constant.

4.6.2 Results

Simulation results are summarized in Table 4.2. Each model was evaluated on the prediction of frequency effects similar to the evaluation of the previous simulation (see Table 4.1 for the frequency classes used in the PSC simulations). However, in addition to the six early fixation measures, we also evaluated the models on the following so-called late measures: regression path duration (RPD, also called go-past duration, the sum of all fixations including previous locations beginning from the first fixation on a word until leaving it to the right), total fixation time (TFT, sum of all fixation on a word), rereading time (RRT, time spent on a word after leaving it and returning to it), first-pass regression probability (FPREG, the probability of regressing from a word in first pass), and the probability of rereading a word after leaving it to the right (reread). Note that first-pass regression probability is not literally a late measure. However, we call it late here because in our model all regressions are caused by late processes. Except for Simulation 2a,

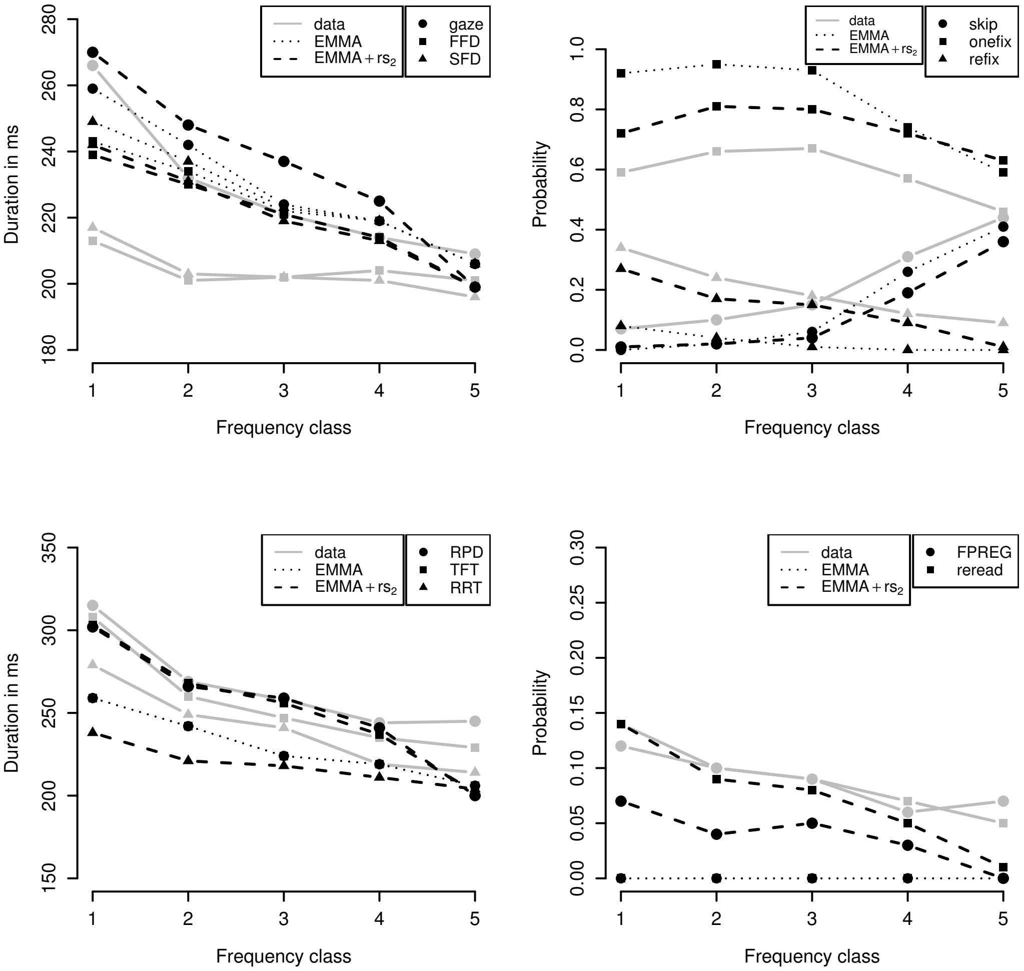

Figure 4.4: Predictions of Model 2 (EMMA, dotted lines) vs. Model 7 (EMMA+rs$_2$, dashed lines) vs. experimental data (grey solid lines) for the Potsdam Sentence Corpus. The figure shows means of early (first row) and late measures (second row) as a function of frequency class. Each row shows reading time durations on the left and probabilities on the right side.

all models were fit and evaluated on the full dataset that contained trials with regressions. Like in Simulation 1, the first and the last word of each sentence were excluded from the analysis. Following the corpus study in Kliegl et al. (2004), we removed words with first fixation durations longer than 1000 ms and words with gaze and total fixation durations greater than 1500 ms from empirical dataset. This reduced the corpus by a number of 79 words. The results shown in the table were obtained by running 100 iterations on the PSC with the respective parameter sets. For each model the best fit was determined in the way described in Simulation 1.

4.6.2.1 PSC vs. SC

Simulation 2 was carried out on the PSC using the pure EMMA model without retrieval or surprisal information. For comparing the model performance on the PSC versus the Schilling Corpus, row 2a in Table 4.2 shows the model performance when trials containing inter-word regressions (40%) were not considered in the analysis. For this case, only early measures were compared. Encoding factor K and T_{prep} were re-estimated. The predictions have a good correlation with the observed frequency effects ($R_{early} = 0.86$). Numerically, the predictions deviate more from the data than for the Schilling Corpus, but the $RMSD$ is still reasonable with a value of 0.326. Note that $RMSD$s are not directly comparable between corpora. $RMSD$s for the PSC are generally a bit lower because the standard deviations used for normalization are higher than in the Schilling Corpus.

4.6.2.2 Influence of Parsing Difficulty

In Table 4.2 the PSC simulations are sorted by goodness of fit as defined by the total correlation R, which is the mean of R_{early} (correlation for the early measures) and R_{late} (correlation for the late measures). It shows that the model performance on predicting frequency effects gradually improves through the extension with measures of surprisal and retrieval. In order to provide a baseline for the EMMA+ models, Simulation 2 was analyzed again on the complete dataset including trials that contained regressions (see row 2b). The results of 2b show that the fit for late measures (R_{late}) is very low, which results in a total R of 0.67. That is expected because three of the late measures (RRT, FPREG and reread) are not predicted at all by the model due to the lack of regressions. Note that although Model 2 did not produce Time Out regressions, some backward saccades happened due to motor error. These did, however, not produce enough data to report mean RRTs over frequency classes: only six words out of 85,000 (850 analyzed corpus words times 100 simulations) were reread.

For the following simulations, the parameters F and P were estimated if the model used retrieval or surprisal, respectively. In Simulation 3 (EMMA+s$_1$), the fit for the early measures improves ($R_{early} = 0.93$), but here still no time outs were produced, as s$_1$ is only modulating word encoding time. In contrast, in Model 4 (EMMA+r), Time Out regressions were produced as a consequence of retrieval difficulty in 18% of the trials. That, of course, improved the prediction of late measures considerably, resulting in an R_{late} of 0.86. Note, however, that R_{early} (0.90) is not as good as with EMMA+s$_1$. Simulation 5 (EMMA+s$_2$) used high-level surprisal that interacts with the model through time outs just like retrieval. Interestingly, it produced a slightly better fit than EMMA+r, especially in early measures ($R_{early} = 0.92$). Combining retrieval and low-level surprisal in Simulation 6 (EMMA+rs$_1$) results in about the same fit as Simulation 5. However, the combination of retrieval and high-level surprisal in Simulation 7 (EMMA+rs$_2$) improves R_{late} even more and results in a total R of 0.91, with a fairly good $RMSD$ of 0.206.

Fig. 4.4 compares the performance of pure EMMA (Simulation 2) with that of the best model, EMMA+rs$_2$ (Simulation 7). In the early probability measures (upper right panel), one can see that EMMA+rs$_2$ produces more refixations, which is also the reason for the prediction that gaze durations are generally longer than first and single fixation durations (upper left panel), which was not quite captured in pure EMMA. The predictions for late duration measures (lower left) show a good fit of TFT and RPD in the complex model up to frequency class 4 with a disproportionate drop in class 5. Also the RRT means are well correlated with the data, whereas the simple model did not predict RRT values at all. It looks similar for late probabilities (lower right): While pure EMMA does not predict any regressions, EMMA+rs$_2$ shows a nearly perfect fit for reading proportions up to frequency class 4 and a little low but well correlated mean proportions of first-pass regressions.

As an additional assessment of the effects of surprisal and memory retrieval, we did a linear regression analysis for selected eye-tracking measures using the predictors log frequency, length, log retrieval, and surprisal. This was done to see which of the six EMMA models predict variance that is explainable by surprisal and retrieval values. Simply reporting means in the same way as it was done for frequency effects would not be informative for surprisal and retrieval as their effects exhibit much interaction with other factors. In order to ensure that the incorporation of surprisal and retrieval information does not just add random or redundant variance to the simulation results, the linear regression models should have sensible estimates for both predictors. This means that, ideally, surprisal effects should be significant in the output of simulations that included surprisal (EMMA+s$_1$, EMMA+s$_2$, EMMA+rs$_1$, and EMMA+rs$_2$), retrieval effects should be significant for EMMA+r, EMMA+rs$_1$, and EMMA+rs$_2$, and none of the two predictors should be significant for the pure EMMA simulation.

We fit linear models on the output of all six EMMA simulations for four selected dependent measures in the statistics software R R Core Team (2012). Equation 4.4 shows the linear model for first-fixation duration (FFD) as for an example:

$$FFD_i = \beta_0 + \beta_1 log(freq_i) + \beta_2 len_i + \beta_3 s_i + \beta_4 log(r_i) + \epsilon_i \tag{4.4}$$

For each predictor, β is the coefficient to be estimated. All predictors were centered at zero.

Fig. 4.5 plots estimates and 95% confidence intervals for surprisal and retrieval. It shows that surprisal and retrieval are significant predictors in almost all EMMA models that incorporate them but not in others, with some exceptions: Surprisal is not significant for FFD in model EMMA+rs$_1$ but is significant in model EMMA+r for RPD and FPREG. It seems that retrieval here subsumes some of the variance that would also be caused by surprisal. Indeed, both predictors are slightly correlated with $r = 0.15$. The fact that surprisal is not significant in model EMMA+s$_1$ for first-pass regressions, on the other hand, is expected, because this model did not produce any regressions. Retrieval estimates are always significant where it would be expected. They are, however, also significant in model EMMA+s$_2$ for RPD and FPREG which, again, points toward a certain correlation with surprisal.

The linear modeling results are in accordance with the results on human data reported in Boston et al. (2011). Boston and colleagues fit linear mixed effects models on the PSC data and reported significantly positive coefficients for both surprisal and retrieval when predicting SFD, FFD, RPD, TFT, and FPREG. Table 4.3 shows surprisal and retrieval coefficients of regression models on the output of EMMA+rs$_2$ and, where available, the corresponding human data as reported in Boston et al. (2011). Note that the coefficients estimated here and those estimated in Boston et al. (2011) are not directly comparable because the linear models used are different. Boston et al. (2011) used more complex linear mixed models including besides surprisal and retrieval word length, word predictability, unigram frequency, and bigram frequency. Item and participant variation were included as random intercepts. Accounting for individual differences is necessary in the case of human data. In our simulations, however, the variance caused by different simulation runs is negligible, which makes the use of mixed models unnecessary. Without accounting for item and participant variation in the human data, however, retrieval effects in particular could not be detected (note the small coefficients for retrieval in the Boston et al. models).

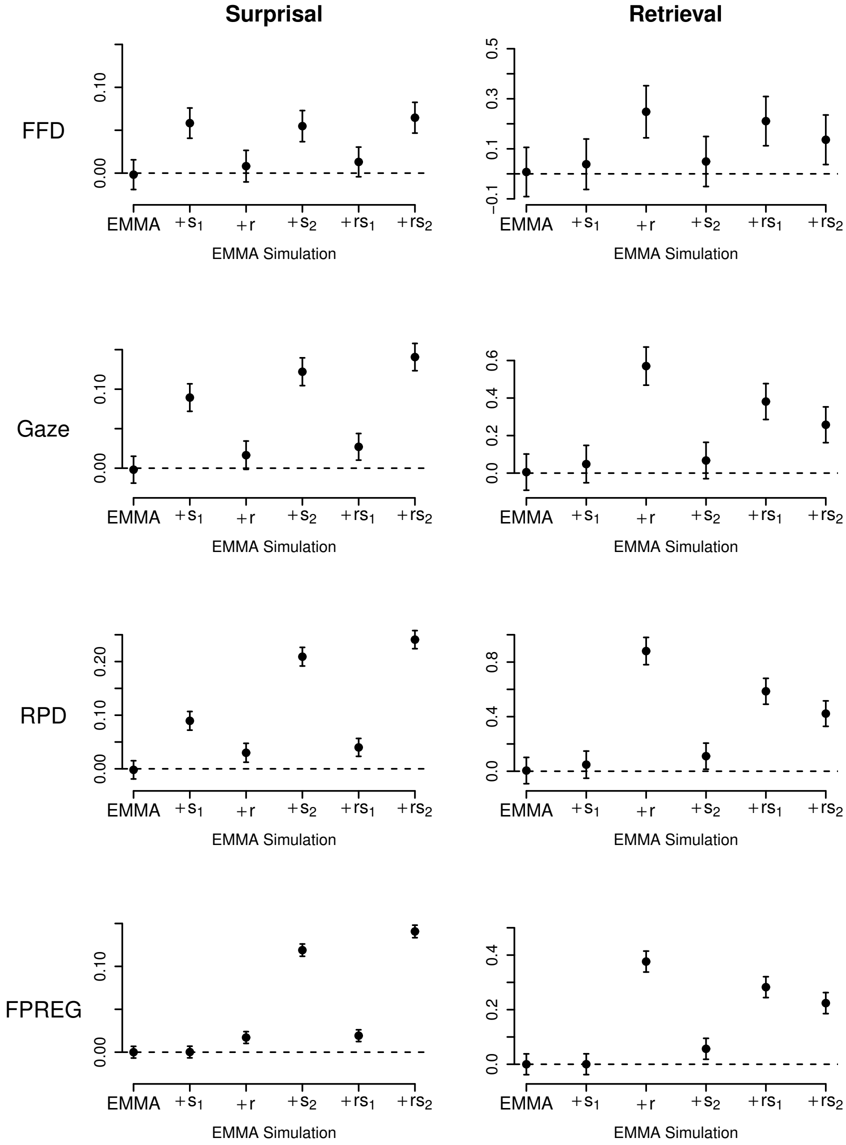

Figure 4.5: Coefficients and 95% confidence intervals for predictors surprisal and retrieval, estimated by linear regression. Predictors were log frequency, length, log retrieval, and surprisal. Coefficients are plotted along the y-axis for surprisal on the left side and retrieval on the right side. Regressions were carried out on the simulated data of all six EMMA models (shown on the x-axis). 95% confidence intervals that do not cross 0 indicate statistical significance at $\alpha = 0.05$.

Table 4.3: Linear regression results for predictors retrieval and surprisal

		Model EMMA+rs$_2$			Data (Boston et al., 2011)		
Measure	Predictor	Coef.	SE	t / z	Coef.	SE	t / z
SFD	Retrieval	0.102	0.056	1.8	0.00015	0.00001	18.2
	Surprisal	0.034	0.013	2.7	0.04384	0.00200	21.9
FFD	Retrieval	0.136	0.051	2.7	0.00016	0.00001	21.1
	Surprisal	0.065	0.009	7.1	0.05209	0.00179	29.0
Gaze	Retrieval	0.258	0.049	5.3			
	Surprisal	0.141	0.009	16.0			
TFT	Retrieval	0.439	0.047	9.4	0.00008	0.00001	8.0
	Surprisal	0.202	0.008	23.8	0.04588	0.00239	19.2
RPD	Retrieval	0.422	0.048	8.9	0.00010	0.00001	9.3
	Surprisal	0.241	0.009	28.0	0.05530	0.00253	21.8
FPREG	Retrieval	0.224	0.020	11.4	0.00026	0.00008	3.5
	Surprisal	0.141	0.004	37.7	0.16890	0.01767	9.6

Note. For FPREG z-values are shown, otherwise t-values. FPREG was modeled with a generalized linear model with a binomial link function for EMMA and a generalized linear mixed model by Boston et al. (2011). For all other dependent measures a linear model was used for EMMA's predictions and a linear mixed model by Boston et al (2011).

4.7 Discussion

The results show that the extension with surprisal and retrieval information considerably improves EMMA's predictions for fixation measures. The interaction of post-lexical processing with EMMA through Time Out regressions enables the model to predict regression-related measures. The best model was EMMA+rs$_2$, which combines retrieval with high-level surprisal, both interacting with EMMA through time outs. Compared to low-level surprisal, the high-level version improves the model much more. The main improvement, however, is due to the possibility of making regressions, which is not possible in EMMA+s$_1$. A fairer comparison between both surprisal versions is between EMMA+rs$_1$ and EMMA+rs$_2$, which both have the ability for Time Out regressions. When we compare each of these two models with EMMA+r, it shows that s$_1$ improves the prediction of both early and late measures a bit and that s$_2$ improves only the prediction of late measures but more so than S$_1$ does. This means that both surprisal versions might be complementary and could be combined in one model. In any case, surprisal, whether high-level or low-level, seems to have more effect on early measures than retrieval when we compare EMMA+s$_1$ and EMMA+s$_2$ with EMMA+r. This is interesting because it is consistent with the results of experimental and corpus studies reported above.

4.8 General Discussion

The primary goal of the current chapter was to make two contributions: First, we replicated the EMMA reading simulation of Salvucci (2001) in a more recent ACT-R environment and extended it with simulations on the German Potsdam Sentence Corpus, thus evaluating EMMA on two different languages. Second, we presented an approach of augmenting EMMA with computational measures of post-lexical processing. The results showed that a combination of retrieval and surprisal substantially improves EMMA's predictions of fixation measures. The implementation of Time Out regressions (Mitchell et al., 2008) in a way similar to E-Z Reader 10 enabled the model to predict regression rates and rereading time. The simulation results corroborate the assumption that retrieval and surprisal are complementary in their influence on eye movements. This can be concluded from the fact that a combination of both predictors results in a better model than using just one of them, and that surprisal has more effect on early measures than retrieval has. The framework's components (ACT-R, EMMA, parser) were chosen with the aim for flexibility and expandability. The simulations presented here were intended as a general demonstration and should serve as a step toward a further precise investigation of the interaction between eye movements and language comprehension. The use of the general modeling architecture ACT-R allows for an easy integration of the model with other sorts of linguistic or psychological factors. Also, all existing simulations that used the cue-based retrieval parsing architecture (e.g., Lewis & Vasishth, 2005; Patil et al., 2015, 2012; Vasishth et al., 2008; Vasishth & Lewis, 2006) can be further investigated by using the published parsing rules seamlessly with the eye movement control model.

4.8.1 Comparison with E-Z Reader

The EMMA/ACT-R model makes some simplifying assumptions with respect to eye movement control and its interaction with parsing. EMMA is a simplified eye movement model, designed for application in various cognitive domains. However, reading is undoubtedly a very specialized and highly trained task that involves enormous complexity. An example of the training aspect is that in E-Z Reader a forward saccade is automatically programmed after a first stage of lexical identification and before the attention shift. In EMMA, saccade programming always starts at the same time as the attention shift and word recognition. As a consequence, most of the word recognition in EMMA happens through preview and often finishes before the eyes have moved to the respective word. For that reason, most Time Out regressions are already initiated when the eyes are still fixating on word w_n (the word with post-lexical difficulty) and therefore target word w_{n-1}. In contrast, regressions triggered by slow integration failure in E-Z Reader would be initiated most of the time from w_{n+1}, at least that seems to be suggested in Reichle et al. (2009). However, this difference might not be a problem for the EMMA model, at least as far as qualitative predictions are concerned. In fact, in the three experiments that are modeled in Reichle et al. the most relevant regression-related predictions are regressions *out* of the target region. In the following, these three experiments shall be briefly described including a short discussion of EMMA's capabilities with respect to according predictions.

The first experiment simulated clause wrap-up effects (Rayner, Kambe, & Duffy, 2000). The critical observations and model predictions for clause-final words were an increased number of refixations and an increased regression probability from these words toward the previous region. In order to predict clause wrap-up effects in EMMA, further assumptions would have to be incorporated into the parsing model, because it does not contain specific processes related to the end of a clause. But, assuming that wrap-up operations increase the length of the integration stage, EMMA would be expected to make the correct predictions. The second experiment was about the effects of plausibility and possibility violations (Warren & McConnell, 2007). Possibility violations are detected early, observed as increased first fixation durations. The effect of implausibility appears later, increasing gaze durations and the probability of regressing out of the target word. As our extension of EMMA concerned only syntactic processing, the model does not predict semantic effects. A hypothetical version of EMMA could include a model of world knowledge similar to Budiu and Anderson (2004) that processes the result of syntactic integration, adding extra time to the integration stage. However, for a process model to account for the time-course difference between plausibility and possibility, the detection of both has to occur in distinct stages. An explanation for the earlier detection of possibility violations might be that such words are highly unexpected (and unfrequent) in the respective context so that predictability or a lexicalized version of surprisal could account for the effect. Assuming surprisal affects word recognition (as in the model EMMA+rs$_1$), it would produce an early effect for possibility violations. Finally, the third experiment discussed in Reichle et al. (2009) can be modeled by EMMA straightforwardly. This experiment examined the effects on disambiguating words in constructions that violate the principles of late closure and minimal attachment (Frazier & Rayner, 1982), so-called garden path sentences. In these sentences, on encountering the disambiguating word, the reader realizes that the syntactic structure built up to that point has to be revised. This again shows up as increased fixation durations and regressions out toward an earlier region. On the disambiguating word, the retrieval parser by Lewis and Vasishth (2005) would perform additional retrievals in order to reattach the ambiguous word to the correct node. This would lengthen the integration stage with the consequence of inflated fixation times and first-pass regressions. However, garden paths that lead to reanalysis are detected very early (effects show up in first fixation duration), which is not predicted by Time Out regressions or slow integration failure. Other than normal retrieval processes, a reanalysis is the consequence of a detection of an integration failure. This motivates the assumption that ongoing integration processes are canceled as soon as the error is detected, similar to the mechanism that Reichle et al. (2009) call *rapid integration failure*. This would predict early effects and first-pass regressions in the disambiguating region in a garden path. Simulations with an implementation of this kind are presented in the following chapter.

4.8.2 Conclusion

The presented simulations are a first step toward more advanced models that specify a concrete link between high-level cognitive processes and eye movements. The simulations show that predictions of parsing models contribute to the explanation of variance in an eye-tracking corpus not only statistically but also in an explicit computational model of eye movement control. Time-out regressions explain a lot of the variance in the data, because short one-word regressions are particularly frequent in reading (Vitu & McConkie, 2000). However, the less frequent but more complex regression patterns such as they occur, for example, in ambiguous sentences are particularly interesting and can tell us a lot about the eye-parser connection (e.g., Frazier & Rayner, 1982; von der Malsburg & Vasishth, 2011, 2013; Meseguer et al., 2002). Important issues are the degree of linguistic information that is used for target selection and whether their function is actually to revisit information or something different (e.g., Booth & Weger, 2013; Inhoff & Weger, 2005; Mitchell et al., 2008; Weger & Inhoff, 2007). With the presented framework, it is possible to examine the individual contributions of memory retrieval and expectation to the behavior at certain points of difficulty and the factors that guide long-range regressions. However, the modeling of regression paths is beyond the scope of this thesis and will have to be devised for future studies.

The next step is to investigate the modeling of concrete examples of parsing difficulty. For the corpus study presented here, we used pre-calculated values for retrieval and surprisal. In the following chapter, the actual parsing architecture of Lewis and Vasishth are used in runtime.

Chapter 5

Toward a Complete Eye-Parser Interface

The previous chapter presented a model of an explicit link between the theory of parsing and memory access by Lewis and Vasishth (2005) and the EMMA model of eye movement control by Salvucci (2001). The model parameters were estimated in an evaluation on the Potsdam Sentence Corpus (Kliegl et al., 2004) using pre-computed values for memory retrieval latency and surprisal from Boston et al. (2011). However, the corpus mainly contained relatively simple, short sentences, which is of little value for testing concrete examples of parsing difficulty. In the current chapter, the resulting parameter estimates will therefore be used to generate new predictions for example sentences from the literature. In doing so, the validity of the model can be evaluated by comparing its predictions to empirical data in response to specific forms of parsing difficulty. The advantage of the model presented here is that it is integrated with the fully specified parser by Lewis and Vasishth (2005), so that it generates predictions in runtime without the need to pre-compute a retrieval metric. For reasons of simplicity, the focus of this chapter is on effects of memory retrieval and not surprisal.

In addition to Interface I: *Time Out*, that was presented in the previous chapter, three further elementary interfaces are proposed. Interface II: *Reanalysis* is an early detection of parsing error resulting in a regression similar to "rapid integration failure" in Reichle et al. (2009); A simulation with Interfaces I and II replicates the results of Staub (2010a), who found effects of memory and expectation in distinct locations of object- vs. subject-relative clauses. Interface III: *Underspecification* aborts a costly attachment alternatively to signaling a time-out depending on the task-relevance of the attachment relation. A simulation illustrates at the example of von der Malsburg and Vasishth (2013) how underspecification results from an interaction of eye movement control with parsing and individual differences in working memory capacity. While

Interfaces I and II are interventions by the parser that interrupt the otherwise autonomous saccade programming, Interface III is rather an intervention in the other direction: In the case of underspecification, the parser is cut off by time pressure imposed by eye movement control. Finally, a possible Interface IV: *Subvocalization* is proposed. As another alternative to Time Out, a word ready for integration could be stored in a phonological memory (Baddeley, 2003; Baddeley & Hitch, 1974) for a very short time until there is free capacity of the sentence processor. This could be used to model spill-over effects of parsing difficulty.

Table 5.1 summarizes relevant parameter values used in the simulations in this chapter. If not stated otherwise, ACT-R and EMMA parameters were kept constant at the values estimated for model 4 “EMMA+r” in the corpus study of Chapter 4 or at default values.

Table 5.1: ACT-R/EMMA parameter values.

Simulation	LF	ANS	MAS	MP	VEF	VEE	SPT
4.6 PSC EMMA+r	0.2	0.15	1.5	1.4	0.002	0.4	0.110
5.1 Relative clauses	0.2	0.15	1.5	1.4	0.002	0.4	0.110
5.2 Underspecification	0.2	0.15	3.5	NIL	0.002	0.4	0.110

Note. LF: latency factor, ANS: activation noise, MAS: maximum associative strength, MP: mismatch penalty, VEF: visual encoding factor, VEE: visual encoding exponent, SPT: saccade preparation time.

5.1 Interface II: Reanalysis

5.1.1 Memory and Expectation in Relative Clauses

An experiment by Staub (2010a) has shown effects of both memory retrieval and expectation at different positions within the same sentence. Staub studied the well-known difference between subject-extracted (SRC) and object-extracted (ORC) relative clauses as in Example (19).

(19) a. The employees that [$_{\mathrm{V}}$ noticed] [$_{\mathrm{NP}}$ the fireman] hurried across the open field.

b. The employees that [$_{\mathrm{NP}}$ the fireman] [$_{\mathrm{V}}$ noticed] hurried across the open field.

With remarkable consistency across languages, it has been found that SRCs are easier to comprehend than ORCs (e.g., Frazier, 1987b; Gibson, Desmet, Grodner, Watson, & Ko, 2005; King & Just, 1991; Kwon, Gordon, Lee, Kluender, & Polinsky, 2010; Schriefers, Friederici, & Kuhn, 1995; Traxler, Morris, & Seely, 2002), with the exception of Mandarin Chinese, where the issue is still under debate (Gibson & Wu, 2013; Hsiao & Gibson, 2003; Jäger, Chen, Li, Lin, & Vasishth, 2015; Lin & Bever, 2006; Vasishth, Chen, Li, & Guo, 2013). The two major explanations of this difference in comprehension difficulty are memory-based and expectation- or frequency-based. Memory-based accounts (Gibson, 1998; Grodner & Gibson, 2005; Lewis & Vasishth, 2005; Lewis et al., 2006; McElree et al., 2003) predict difficulty in ORCs due to the increased distance between the embedded verb *noticed* and its subject *employees* or memory interference with the intervening noun *fireman*.

According to expectation-based explanations (Gennari & MacDonald, 2009; Hale, 2001; Levy, 2008; Mitchell, Cuetos, Corley, & Brysbaert, 1995), SRCs are easier to process because their structure is more frequent and more regular than that of ORCs. For most languages, both accounts predict a preference for the SRC. However, for Chinese Mandarin relative clauses the two accounts make opposing predictions, which might be the cause for the long-lasting debate about which construction is preferred: Expectation predicts a subject-relative preference here, too, because SRCs are also more frequent in Mandarin. But as Mandarin is an SVO language and its RCs are pre-nominal, the dependency is more distant in the SRC than in the ORC. Hence, a preference for ORCs is predicted by memory-based accounts in Mandarin. This example shows the importance of generating detailed predictions in order to disentangle the contributions of the two complementary sources of difficulty, memory and expectation.

In their eye-tracking experiment, Staub (2010a) found increased difficulty at two positions in the ORC (19b): At the embedded noun phrase *the fireman*, they found an increased probability of outgoing first-pass regressions (FPRP) and at the embedded verb *noticed* they found increased reading times in gaze duration, first-fixation duration (FFD), and go-past time (regression-path duration, RPD). Staub (2010a) interpreted the results as evidence for both memory-based and expectation-based explanations for difficulty in ORCs. The difficulty on *the fireman* can be explained by expectation: On seeing the embedded subject noun phrase of the ORC, the expectation for an SRC is violated, which would cause surprise. An effect at *noticed* is predicted according to memory-based accounts due to difficulty integrating the distant dependency with the subject. This analysis was supported by the fact that the two observed effects were qualitatively different: Elevated reading times at the ORC verb are compatible with memory-based explanations that predict an increased memory access latency for distant dependencies. More frequent regressions from the ORC subject indicate surprise, causing the reader to interrupt reading and potentially revisit previously read material that lead to the now falsified expectation.

The Staub (2010a) experiment is a good test case for the model developed in the previous chapter for two reasons: First, relative clauses are possibly the best studied constructions in psycholinguistics and the subject-relative preference is an extremely reliable result. Second, the model not only contains a memory component but also an incremental serial parsing mechanism that commits to one structure at a time. This implicates as a sort of expectation of what is coming next by pre-building the necessary structure, which has to be revised when the parser is garden-pathed. Garden-pathing is what happens in the case of encountering an ORC structure when an SRC is expected.[1] Therefore, a natural next step is to implement a mechanism that

[1] Ranked parallel parsers do not assume a revise-on-garden-path mechanism in the sense of Frazier and Rayner (1982) but react by re-ranking possible parses. While this predicts increased difficulty at the same region as predicted by a serial parser, it might have consequences for the prediction of eye movement reactions:

interfaces moments of garden-path with eye movement behavior. This is pursued in the next subsection.

5.1.2 Simulation

When integrating a word that is ambiguous in its function in the current parse, an incremental serial parser commits to one possible continuation. By doing so, the parser creates a kind of *expectation* for subsequent words to conform with this commitment. If the upcoming material cannot coherently be integrated, the earlier decision has to be reanalyzed. When encountering a relative pronoun, the Lewis and Vasishth (2005) parser acts according to a subject preference, always creating an SRC structure. When finding an NP to follow the relative pronoun (the subject NP of an ORC), the parsing rules execute a *reanalysis* of the pre-built SRC structure as a *gapped* ORC structure. In particular, this means that the relative pronoun is made available as a filler that can be retrieved as the object at the ORC verb. For that, an additional retrieval of the relativizer DP has to be executed.

This revision process, which occurs at the ORC subject NP, is compatible with the assumptions of expectation-based theories, namely that at this point surprise occurs and expectations have to be revised. The expectation in this case is the syntactic structure built so far, which has to be altered in order to consistently incorporate the present input.

With the Time Out extension, presented in the previous chapter, elevated reading times and occasional regressions would be predicted at the ORC subject, because the additional rule firing and retrieval of the revision process delays the completion of the integration at this point. This would be the case if a "slow integration failure", as Reichle et al. (2009) call it, would be assumed. However, the case of the SRC/ORC revision is better described as what Reichle et al. (2009) call "fast integration failure". In contrast to slow failure, where simply the integration time takes longer than usual, at a revision, an error is encountered immediately, with the result that previous material has to be revisited. According to the definition of Reichle et al. (2009)'s fast failure, an immediate attention shift is triggered toward the potential cause of the error. In the SRC/ORC revision, the cause of the error (the violation of expectation) is the structural decision made at the relative pronoun. It is reasonable to assume an attention shift toward previous material in the case of a revision but not in the case of a regular retrieval, because, in a revision, previously attached material changes its role, i.e., its position in the sentence structure.

Interface II: Reanalysis is defined as follows:

Interface II: Reanalysis Any rule that changes the attachment of a previously created syntactic object in memory triggers an immediate attention shift toward a point in the sentence that is related to the object in question.

As multiple parses are entertained in parallel, it is not necessary to revise previous commitments and, thus, inflated fixations times rather than regressions might be expected.

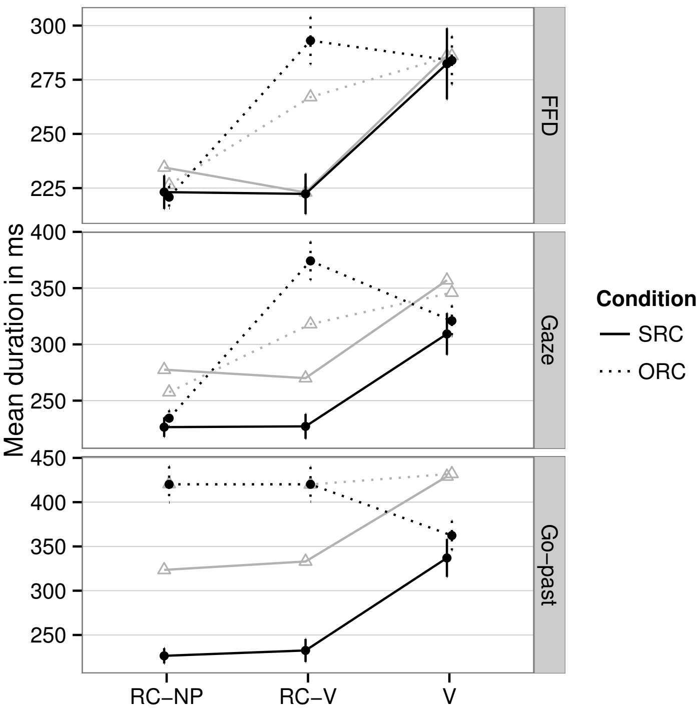

Figure 5.1: Model predictions for reading times in subject- and object-relative clauses at embedded NP, embedded verb, and main verb. Data of Staub (2010) shown in gray. Error bars represent 95% confidence intervals.

The specific target of the attention shift and the potentially accompanying regression is subject to further research. As discussed in the introduction, it is still a difficult question how regression paths behave in reanalysis. For the current simulations, the model is restricted to specifying the source of the regression and not the target. A least-commitment implementation was chosen that selects any target to the left of the current fixation. A possible implementation of target selection in reanalysis will be discussed in the Outlook section of Chapter 6.

5.1.3 Results

Simulations were performed using the parameter values that have been estimated with the Time Out model on the Potsdam Sentence corpus in Chapter 4. Each sentence of Example (19) was run 200 times.

Figure 5.1 shows mean reading times for first-fixation duration, gaze duration, and go-past time in comparison with the data (in gray). In first fixations and gaze durations, the Staub (2010a) data shows reading time differences only on the embedded verb, the ORC being slower than the

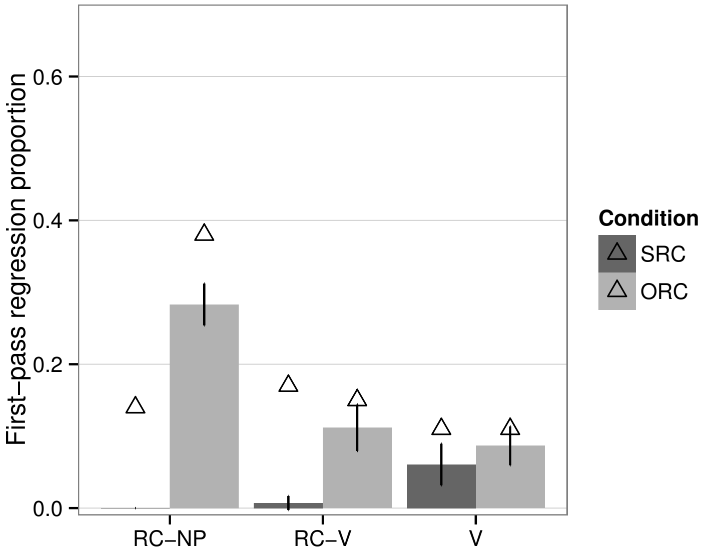

Figure 5.2: Model predictions for first-pass regressions in subject- and object-relative clauses at embedded NP, embedded verb, and main verb. Data of Staub (2010) shown as triangles. Error bars represent 95% confidence intervals.

SRC. In go-past times, there is a difference also on the relative clause NP, because there are more regressions occurring at this point in the ORC.

Qualitatively, the model predicts the data exactly in all three eye movement measures. Numerically, there is a remarkably close fit in first-fixation durations. Also in gaze and go-past times, the predictions are numerically in a similar range as the data, but the SRC reading times are underestimated in the predictions by 50 to 100 milliseconds.

First-pass regression proportions are shown in Figure 5.2. The empirical means are indicated by the black triangles. They show an increased regression rate in ORCs at the relative clause NP. Like for reading times, the model predictions for regressions are very close to the data. However, no first-pass regressions are predicted for the SRC at the NP or the verb inside the relative clause. The model predicts the major empirical finding of an increased regression rate at the relative clause NP for the ORC. It does, however, additionally predict a slight increase at the ORC verb, which has not been found in the data.

5.1.4 Discussion

The simulation tested the model developed in Chapter 4 on specific predictions for English relative clauses. No parameter fitting to the data was performed. The model with parameters as estimated on the Potsdam Sentence Corpus generated predictions that are qualitatively and quantitatively very close to the data. The Time Out interface developed in the previous chapter predicts memory-based differences and magnitudes in three different reading time measures in three regions remarkably well. This shows that the parameter estimation that was done on the Potsdam Sentence Corpus of German is representative and holds for experimentally controlled comparisons of memory-related complexities also in English. The newly introduced Reanalysis interface II, which is similar to Reichle et al. (2009)'s rapid integration failure, correctly predicts increased first-pass regressions due to an invalidated prediction.

Taken together, the simulation results confirm Staub's assumption that the qualitatively distinct effects on the RC noun phrase and the RC verb indicate distinct sources of difficulty, namely expectation and memory, respectively. Note that the mechanism of interaction in the model is in both cases the same: The parser intervenes in the forward movement of the eyes, triggering a regression. However, the timing differs in that the intervention happens earlier for violated expectations: While memory-induced time-outs are triggered after recognition of the next word, reanalysis regressions are triggered as soon as the parser detects that the input is unexpected. Due to this timing difference, the time-out regressions on the verb are mostly canceled before they are executed, because the memory processes are not delayed long enough before normal reading is resumed. The planning and canceling of regressions thus appears as inflated reading times. Reanalysis, in contrast, is triggered earlier and the parser has to perform additional actions for revising structure, which leaves enough time for the regression to be completed. This is an example of a simple mechanism producing complex behavior: The same mechanism — triggering a regression — under certain circumstances produces qualitatively different effects based on relative timing between parsing and eye movement control.

The model predictions capture the major findings of Staub's study. However, the model is very simplifying and consequently deviates in its predictions in some ways from the data. First, a subject-relative preference is hard-coded into the model, whereas human readers might also expect ORCs in some cases. The underestimation of SRC gaze and go-past durations in the model would be less significant with a probabilistic decision between pursuing an SRC or ORC structure.[2]

Second, first-pass regression proportions are predicted to be slightly increased on the RC verb in the ORC, although the only empirical effect reported was on the RC noun phrase. This indicates that not all of the time-out regressions that are triggered at this region were canceled but some had enough time to be executed, meaning that the model predicts memory-based regressions on the verb which are not supported by Staub's data. The biggest effect in the predictions of first-pass regressions is, however, on the ORC subject, consistent with the data.

[2] In ACT-R, it is possible to learn the utilities of parsing rules over a number of trials, based on the number of successful applications. It would be necessary to simulate reading of a corpus that represents the natural distributions of relative clauses and similar structures.

Third, expectation-based first-pass regressions are only predicted on the determiner of the ORC subject but not on the noun. In the data, however, the effect is found at the determiner *and* the noun (this is not plotted here but will be shown in Section 5.3). This is most likely a spill-over effect due to delayed parsing in some trials, which is not predicted in this case by the model. It is possible in the model that parsing processes on word n influence the fixation time on word n+1: This happens when a time-out is initiated while the eyes have already moved on to word n+1. However, this does not happen in this particular simulation, because the detection of the validated expectation is instantaneous, meaning it is part of the first parsing production firing at the determiner. Hence, reanalysis is always initiated before the time-out production can fire. This might be a technical flaw in the model which will have to be addressed in the future. However, an alternative cause for spill-over effects — especially, if they span multiple words — may be that readers delay parsing in order to collect more information before making structural decisions. A likely mechanism for delaying parsing processes in this way is to hold words temporarily in the articulatory loop (Baddeley, 2003; Baddeley & Hitch, 1974). This possibility will be discussed and tested with a brief simulation in Section 5.3 of this Chapter.

5.2 Interface III: Underspecification

5.2.1 Good-Enough Parsing

The good-enough approach to sentence processing (Ferreira et al., 2002; Sanford & Sturt, 2002) suggests that readers strategically adapt their efforts to task demands with the consequence of sometimes not arriving at a complete parse. This implicates that readers leave some structural relations underspecified in order to save processing time if the effort does not seem necessary for the task at hand.

Recognized as an example of underspecification is the finding that some ambiguous attachment relations are read faster than their unambiguous counterparts. For example, Traxler et al. (1998) and Traxler (2007) studied sentences like (20c) that were globally ambiguous with regard to the attachment of the relative clause to one of the noun phrases *sister* or *writer* and compared them to sentences where the relative clause was unambiguously attached *high* (20a) or *low* (20b).

(20) a. The writer of the letter/ that had/ blonde hair/ arrived this/ morning.

b. The letter of the writer/ that had/ blonde hair/ arrived this/ morning.

c. The sister of the writer/ that had/ blonde hair/ arrived this/ morning.

Both studies found an ambiguity advantage at the disambiguation region *blonde hair*. An analysis of individual working memory capacity in Traxler (2007) revealed that high-capacity readers showed an expected preference for high attachment (NP1). Low-capacity readers, in contrast, showed no such preference. According to Traxler, it might be that low-capacity readers leave the attachment underspecified or the selection of the attachment site is the product of a balance between a high-attachment preference and recency.

Kemper et al. (2004) studied a main clause/relative clause ambiguity such as (21).

(21) a. The experienced soldiers/ warned about the dangers/ before the midnight raid.

b. The experienced soldiers/ warned about the dangers/ conducted the midnight raid.

c. The experienced soldiers/ spoke about the dangers/ before the midnight raid.

d. The experienced soldiers/ who were told about the/ dangers conducted the midnight raid.

In (21a) and (21b), the role of *warned* is temporarily ambiguous between the main verb and the embedded verb of a reduced relative clause. In (21b), the ambiguity is resolved toward the non-preferred reduced-relative reading. In (21c), the verb *spoke* unambiguously induces a main verb reading. Kemper and colleagues found the expected difficulty at *conducted* in the non-preferred condition (21b). However, in contrast to the studies by Traxler and colleagues, they found no ambiguity advantage. An analysis of working memory differences showed that low-capacity readers had more difficulty resolving the ambiguity, which was indicated by slower reading and higher regression rates at the disambiguating region.

A possible account of when attachments are underpecified and thus an explanation for the difference between Traxler's experiments and Kemper et al. (2004) is offered by *construal* (Carreiras & Clifton, 1993; Frazier & Clifton, 1997). This theory differentiates between "primary" and "nonprimary" relations, where primary roughly stands for relations that are obligatory for deriving a coherent message (e.g., verbs and their arguments). The attachment of primary relations is always carried out according to garden-path theory (Frazier, 1987a) while the definite attachment of nonprimary relations can be suspended by loosely associating it with the last theta domain (Frazier & Clifton, 1997).

More evidence for construal and the good-enough account (Ferreira et al., 2002; Sanford & Sturt, 2002) comes from a self-paced reading study by Swets et al. (2008). Similar to Traxler et al. (1998) and Traxler (2007), Swets and colleagues studied ambiguous relative clause attachments like in Example (22).

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

b. The son of the princess who scratched himself in public was terribly humiliated.

c. The son of the princess who scratched herself in public was terribly humiliated.

Swets et al. manipulated task demands by using either superficial comprehension questions or questions that specifically queried the interpretation of the relative clause. The results showed an ambiguity advantage in (22a) vs. (b) and (c) at the disambiguating reflexive only when questions were superficial, indicating that question type affected the readers preference to leave the RC attachment underspecified. In addition, in the condition with questions targeting the relative clause (RC question condition), question response times were elevated for ambiguous sentences. This indicates that even in the RC question condition the attachment was sometimes left unspecified and had to be resolved during the question answering phase. In the RC question condition, a disambiguation toward NP1 (22b) resulted in longest reading times, pointing to a preference toward NP2 in the initial attachment, which had to be revised at the reflexive. For a detailed analysis of the Swets et al. data and the compatibility with assumptions about underspecification, see Logačev and Vasishth (2014).

Finally, von der Malsburg and Vasishth (2013) presented clearer evidence for the influence of working memory capacity on the preferences for underspecification. They conducted an eye-tracking experiment using the stimuli of a Spanish study by Meseguer et al. (2002) and analyzed the participants' individual scanpaths (von der Malsburg & Vasishth, 2011) and working memory capacity. In the sentences studied (Example 23), an adjunct (*cuando los directores...*) was temporarily ambiguous between modifying the main verb *dijo* as a temporal adverbial clause (high attachment) or the embedded verb *se levantaran* as a conditional (low attachment).

(23) El profesor dijo que los alumnos se levantaran del asiento

a. cuando los directores **entraron** en la clase de múusica. (HIGH)

b. cuando los directores **entraran** en la clase de música. (LOW)

c. **si** los directores entraban en la clase de música. (UNAMB)

"The teacher said that the students had to stand up from their seats when/when/**if** the directors **came**/**come**/come into the music class."

The attachment site was disambiguated at the verb *entraron* (indicative)/*entraran* (subjunctive) toward HIGH (23a) or LOW (23b) attachment, respectively. In the unambiguous (UNAMB) condition (23c), using the word *si* ("if") instead of *cuando* ("when"/"if") unambiguously signaled LOW attachment of the adjunct as a conditional.

Von der Malsburg and Vasishth found an ambiguity advantage in the pre-verbal region *cuando los directores*: First-pass reading times were faster in ambiguous conditions. This effect was driven mainly by low-capacity readers (see Figure 5.3). This finding corroborates the assumption of Traxler (2007) that low-capacity readers underspecify more often. More supporting evidence for this assumption comes from the proportions of rereading of the whole sentence after seeing the disambiguating region (see Figure 5.4). For high-capacity readers, the rereading proportion was highest for condition (a), where the disambiguation was toward high attachment. In contrast, low-capacity readers showed no difference in rereading proportions between conditions. According to von der Malsburg and Vasishth, this indicates that high-capacity readers complete the

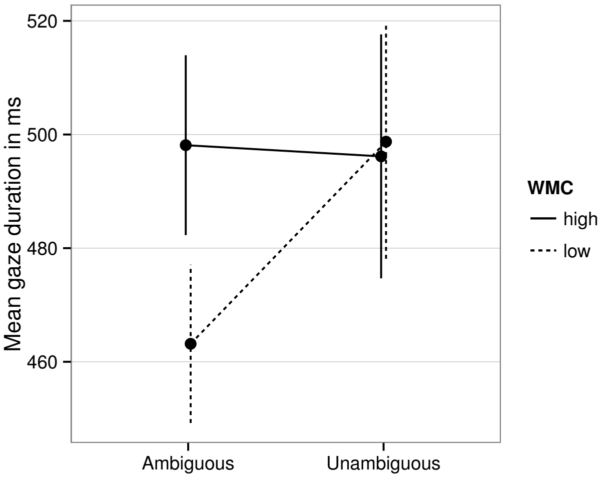

Figure 5.3: Ambiguity advantage for low-capacity readers in the data of von der Malsburg and Vasishth (2013). Shown are gaze durations in the pre-verbal region (*cuando/si los directores*) for ambiguous (a and b) and unambiguous (c) conditions, grouped by high and low working memory capacity. Error bars represent 95% confidence intervals.

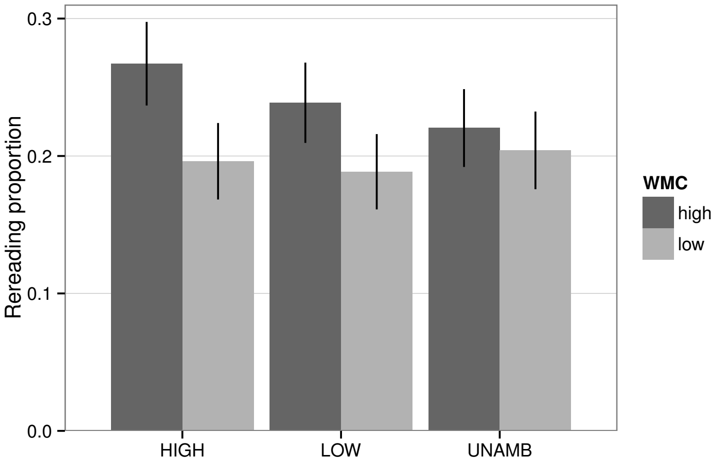

Figure 5.4: Proportions of sentence rereading in the Data of von der Malsburg and Vasishth (2013) for ambiguous high (a), low (b), and unambiguous (c) conditions, grouped by high and low working memory capacity. Error bars represent 95% confidence intervals.

attachment of the adjunct more often and thus have to reanalyze later, whereas low-capacity readers have no need for reanalysis because they mostly leave the attachment unspecified in the ambiguous conditions. The reason for reanalysis occurring predominantly in the high-attachment condition is that the conditional interpretation with a low attachment is initially preferred with the word *cuando*.

The evidence summarized above suggests that ambiguous attachment relations are strategically underspecified and that readers with low working memory capacity leave an attachment underspecified more often than do high-capacity readers. It is not clear, however, how this adaptation works. Is it necessary to assume that low-capacity readers use a different parsing strategy, or can the difference be explained by a common mechanism? So far there is no detailed model of the good-enough account that could clarify this issue.

Here, it is proposed that underspecification results from a common strategy that aims for uninterrupted reading whenever possible. In particular, for nonprimary relations in the sense of Frazier and Clifton (1997), an attachment is treated as non-obligatory and completed only if enough time is left before the next word is ready to be integrated. Thus, underspecification is not a deterministic process but dynamically adapts to the relative timing of the attachment process and autonomous low-level eye movement processes. If an attachment is easy and proceeds fast in relation to the "default" reading speed (in the sense of *uninterrupted* reading), the attachment is completed. If, however, the attachment could not be made without interrupting the progress of the eyes, it is abandoned in a trade-off with reading speed. An influence of working memory differences is predicted by the assumption that low-capacity readers on average take longer to complete the attachment than high-capacity readers, resulting in more cancellations that leave the attachment underspecified.

5.2.2 Simulation

The model was implemented that defines *good-enough* parsing as the result of an interaction between the parser and eye movement control as explained above. For non-obligatory relations, the time to attach is constrained by the time needed by saccade programming and low-level processes to identify the next word:

Interface III: Underspecification For relations with low utility, an attachment attempt is aborted as soon as the next word is ready for integration, so that reading proceeds uninterrupted.

This implementation naturally predicts an influence of working memory capacity if it is defined as a measure of speed and accuracy of retrieval processes such as goal buffer source activation W in ACT-R (see Chapter 2). Working memory capacity has been modeled in this way before: Daily, Lovett, and Reder (2001) modeled individual differences in the digit span task (Lovett, Reder,

& Lebiere, 1999) in ACT-R by manipulating the source activation W for the goal buffer. In the context of language comprehension, van Rij, van Rijn, and Hendriks (2013) used this method by manipulating W for modeling individual differences in pronoun interpretation. The amount of activation W is equally distributed between sources j in the goal buffer, i.e., the chunks that spread activation to related memory items. Thus, the value of W defines how strongly relevant information from the goal buffer is used for memory retrieval. Therefore, higher values of W improve speed and accuracy of retrieval processes. In this way, the term working memory capacity is defined as a measure of how well an individual is able to separate information in memory that is relevant to the current task from currently irrelevant information — a kind of focusing of cognitive attention.

For an exemplary simulation, the Lewis and Vasishth (2005) model was extended with parsing rules for sentence constructions as used by von der Malsburg and Vasishth (2013) and defined the adjunct attachment as non-obligatory. In particular, when attaching the adjunct, the parser creates the structure for the adjunct clause and then signals that integration is complete, so no time-out rule will fire. It then attempts to retrieve both potential attachment sites. This process is faster and more accurate for high-capacity readers, because, with a high source activation W, the retrieval cues activate the correct retrieval targets more strongly. Thus, high-capacity readers are predicted by the model to underspecify less often than low-capacity readers.

The mechanism, however, is the same for all readers: As soon as the next word is ready for integration, an ongoing attachment process is abandoned. Following Swets et al. (2008), an abandoned attachment is not corrected later in the sentence. If an attachment was made, however, contradicting disambiguation information leads to a repair operation triggering a regression toward the beginning of the sentence.

60 participants were simulated reading the three attachment conditions of von der Malsburg and Vasishth (2013), HIGH, LOW, and UNAMB, 20 times each. EMMA parameters and the latency factor were left at values estimated in the Potsdam Sentence Corpus evaluation in Chapter 4. Other parameters were set to the values used in Lewis and Vasishth (2005). Two parameters were changed for this simulation, however: *Mismatch penalty* MP was set to `NIL` to switch of partial matching in order to reduce interference and misretrievals as this was not relevant here; the *maximum associative strength* parameter MAS had to be increased from 1.5 to 3.5 in order to have enough spreading activation from the goal buffer for differences in W to have an effect. For simulating individual differences in working memory capacity, the method of Daily et al. (2001) was used to randomly assign to W a value drawn from a normal distribution with mean = 1.0 and sd = 0.25.

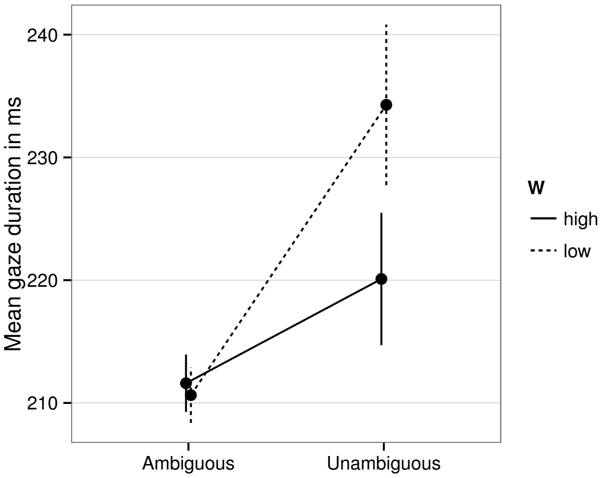

Figure 5.5: Model predictions for gaze durations at *cuando/si* for ambiguous (a and b) and unambiguous (c) conditions, grouped by high and low goal buffer source activation W. Error bars represent 95% confidence intervals.

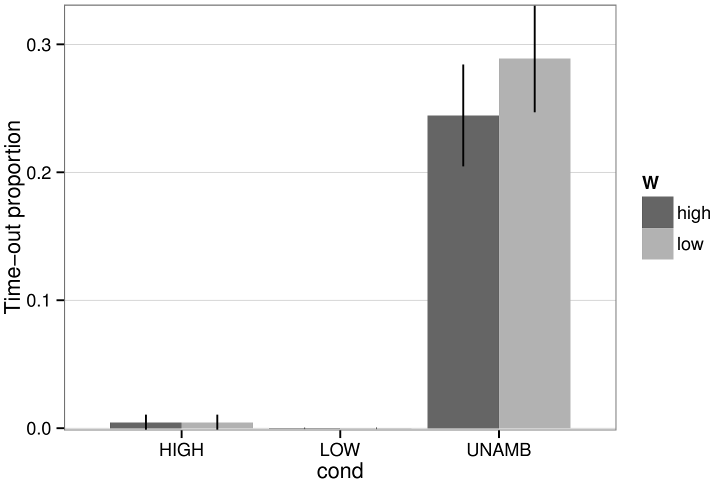

Figure 5.6: Predicted time-out proportions for ambiguous high (a), low (b), and unambiguous (c) conditions, grouped by high and low goal buffer source activation W. Error bars represent 95% confidence intervals.

5.2.3 Results

Simulated participants were grouped into high and low capacity with respect to the randomly assigned goal source activation W by using a median split (the median of W was 0.96). This resulted in 30 simulated high-capacity subjects with mean $W_{high} = 1.19$ and 30 simulated low-capacity subjects with mean $W_{low} = 0.75$. Although no parameter fitting to empirical data was performed, the modeling results are compared to the findings reported in von der Malsburg and Vasishth (2013).

Model predictions for gaze durations for the potentially ambiguous region *cuando/si*[3] are plotted in Figure 5.5. The predictions clearly show an ambiguity advantage which is more pronounced for low-W simulations than for high-W simulations. This is in line with the findings of an ambiguity advantage by von der Malsburg and Vasishth (2013) and others. However, the detailed pattern in the data of von der Malsburg and colleagues (Figure 5.3) is different. While in the simulation, low-capacity readers are generally slower than high-capacity readers, this is not the case in the empirical data. The data shows both groups to be equally fast in unambiguous conditions.[4]

The timing difference between low- and high-capacity readers in the unambiguous condition predicted by the model is due to different time-out proportions in the attachment region, as shown in Figure 5.6. Because attachment is generally slower for low-capacity readers, more time-outs (interface I) are necessary that make the eyes wait for the completion of integration. No time-outs are predicted in the ambiguous conditions, since attachment happens as a process of minor importance in this case.

Figure 5.7 shows the proportions of rereading the sentence after seeing the disambiguating verb *entraron/entraran* in the ambiguous conditions or *enraban* in the unambiguous condition. The model predicts rereading almost exclusively in the ambiguous high-attachment condition, because the parser mostly attaches low and reanalysis is therefore mainly needed when the sentence is disambiguated toward high attachment. A second prediction is that high-capacity readers reread in that condition more often than low-capacity readers. The reason for this is that simulated high-capacity subjects attach more often than low-capacity subjects, as is shown in Figure 5.8. This is the case in both ambiguous conditions and is responsible for the ambiguity advantage seen for both groups. Rereading is, however, only affected in the ambiguous HIGH-disambiguation condition, since only in this case reanalysis is necessary.

Comparing the predictions in Figure 5.7 with the data in Figure 5.4 shows differences. While the rough pattern — most rereading is shown by high-capacity readers in the ambiguous HIGH condition — is consistent, the predictions show that rereading proportions of low-capacity readers

[3] Predictions are shown for one word only and not for the whole pre-verbal region, because the model currently does not predict any spill-over due to delayed parsing processes. Therefore, the effect of underspecification appears immediately at the critical word. See also the General Discussion on this point.

[4] Through personal communication, we know that the empirical pattern shown in Figure 5.3 as found by von der Malsburg and Vasishth (2013) might be questionable. The experiment is therefore currently being replicated by von der Malsburg and colleagues.

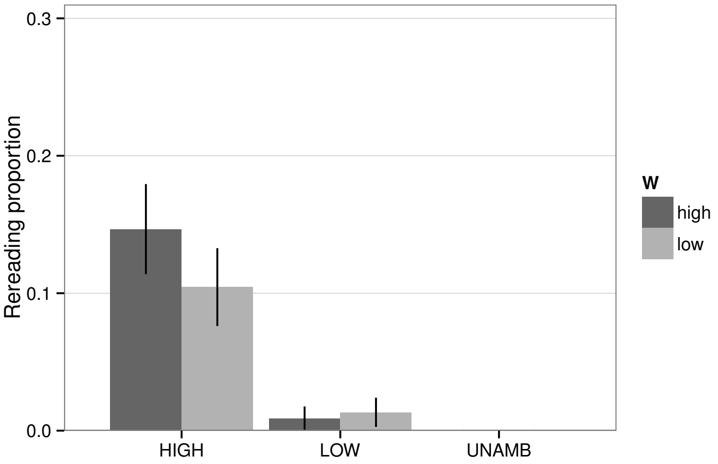

Figure 5.7: Predicted proportions of sentence rereading for ambiguous high (a), low (b), and unambiguous (c) conditions, grouped by high and low goal buffer source activation W. Error bars represent 95% confidence intervals.

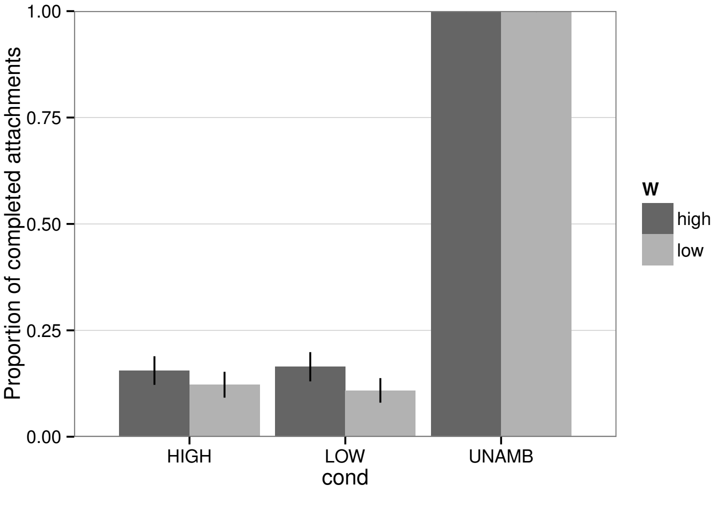

Figure 5.8: Predicted attachment proportions for ambiguous high (a), low (b), and unambiguous (c) conditions, grouped by high and low goal buffer source activation W. Error bars represent 95% confidence intervals.

are also affected by condition, which is not the case in the data. Another difference is that, in the data, there is some rereading in every condition, while the predictions show rereading only in the reanalysis condition (a). Both differences are due to the simplified nature of the model, which will be elaborated in the discussion.

5.2.4 Discussion

Our simple model of good-enough parsing predicts the essential observations: (i) An ambiguity advantage appears at the point of attachment, (ii) low-capacity readers leave an attachment underspecified more often than high-capacity readers, and (iii) high-capacity readers consequently have to reanalyze more often at the point of disambiguation. Most importantly, these predictions are not the result of different strategies for different groups of working memory capacity, but these different behaviors emerge from one common strategy that is an adaptive trade-off between attachment accuracy and reading speed, caused simply by the timing of low-level word identification and oculomotor processes. This simple mechanism leads to an adaptive interaction between parsing, eye movement control, and working memory capacity that predicts the observations mentioned above.

The model used for the above simulation was deliberately kept simple, because the aim was to transparently track predictions to underlying mechanisms. This simplification is responsible for some differences between the model predictions and the empirical observations of von der Malsburg and Vasishth (2013). In the following, four simplifying assumptions in the current model and their implications will be discussed.

Firstly, the model deterministically preferred low attachment (except a small amount of erroneous retrieval of the high attachment site). This explains why rereading is only predicted in the ambiguous HIGH condition (a). In a more realistic model, a utility value for the attachment productions would simulate a non-deterministic preference that would lead to rereading in both ambiguous conditions. However, there is certainly some amount of rereading that is due to other factors than disambiguation such as misreadings, erroneous attachments, or other difficulties due to the length and complexity of the sentence. This is clear from the fact that the data also shows rereading in the unambiguous conditions where no reanalysis is necessary per se.

Secondly, the model deterministically completed the attachment in one hundred percent of the simulation runs in the unambiguous condition. In this condition, the model predicted low-capacity readers to be slower than high-capacity readers, because attachment takes longer for the former. For simplicity, underspecification was only allowed in the model for the specific relation of adjunct attachment. In a more sophisticated model, the trade-off between time-out (the eyes wait for the parser) and attachment underspecification (the eyes cut off the parser) would be defined by a utility value which is learned for different attachment relations by reading experience. Thus, also unambiguous and obligatory relations would have some possibility for being underspecified. A possible model would be the following: For every relation, there is a utility value that decides between the two possibilities time-out (completing the attachment) and cut-off (underspecification). These values are adjusted after every sentence by a positive or negative *reward* according to a task such as answering comprehension questions. An incorrect response to the task will shift the utility toward more accuracy making the completion of attachments more important. A high number of time-outs during the sentence reading will, however, shift the utility toward speed, reducing the importance of attachments, leading to more underspecified relations. The utility learning module of ACT-R would be suitable for this kind of model. In ACT-R, when a reward is triggered, the utility values U of all productions that fired since the last reward are updated with emphasis on more recent productions as defined in Equation (5.1):

$$U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)] \tag{5.1}$$

where α is the learning rate set by a parameter, $R_i(n)$ is the reward value given to production i at time n.

This mechanism would ensure a balance between speed and accuracy according to individual differences and task demands. For low capacity readers, the utility for attachment would automatically turn out generally lower and thus compensate for slower memory processes. As a result, reading speed of low-capacity readers would be more or less the same as of high-capacity readers. This adaptive, yet simple mechanism of global good-enough processing could explain why there is no empirical evidence so far that low-capacity readers generally read slower than high-capacity readers. In addition, a model like this would predict the results of Swets et al. (2008) who found indications for underspecification only in sessions when superficial comprehension questions were asked. The adaptation of utility values according to response accuracy would lead to more attachments for comprehension questions targeting the ambiguous relation and more underspecification for easy questions. This adaptation would occur over time and predict trial effects.

A third simplifying model assumption regards the generation of differences in working memory capacity. While the data shows qualitative differences in the ambiguity advantage (there is no advantage for high-capacity readers) and rereading proportions (there are no differences by condition for low-capacity readers), the model rather predicts quantitative differences: The ambiguity advantage is more pronounced for low-capacity readers and differences in rereading proportions are more pronounced for high-capacity readers. This is a result of the choice of variance between subjects (W being a normal distribution around 1 with sd = 0.25). Choosing a greater variance would increase the differences between capacity groups and potentially predict the qualitative differences found in the empirical data. In order to model the results of von der Malsburg and Vasishth (2013), the variance for W would have to be according to the empirical variance in scores of the reading-span test that assesses individual working memory capacity.

The fourth and last point is an immediacy of effects, i.e., effects are mostly predicted at the word that causes them. What is needed is a mechanism that allows for somewhat delayed parsing processes and predicts spill-over as observed, e.g., in the expectation-based first-pass regressions in Staub (2010a) or the ambiguity advantage in gaze durations across the pre-verbal region in von der Malsburg and Vasishth (2013). A possible mechanism of this kind, using subvocalization, is discussed in the following section.

In summary, it has been shown that a simple common mechanism leads to an interaction between parsing, eye movements, and individual differences predicting observations that are attributed to good-enough processing. Further, a more sophisticated model is proposed that uses adaptive utilities that predict an interaction of underspecifications with task demands.

5.3 Interface IV: Subvocalization

As discussed above, the model presented so far lacks an account of spill-over effects that stretch beyond the time of identification of the next word. The Time Out interface ensures that the integration of word n completes before integration of n+1, so the eyes will progress at most one word ahead before a time-out regression is initiated. Spill-over effects are common, however, and often seen across more than one word. This could be the result of holding words in a short-term storage. There is evidence that subvocalization could play such a role in reading (Baddeley, 1979; Baddeley, Eldridge, & Lewis, 1981; Daneman & Newson, 1992; Eiter & Inhoff, 2010; Kleiman, 1975; Slowiaczek & Clifton Jr., 1980).

An example for spill-over of parsing effects was discussed in Section 5.1 of this chapter. In Staub (2010a), an effect of a disconfirmed expectation of a subject-relative clause manifested itself as inflated first-pass regression proportions on the determiner and the noun of the embedded subject in an object-relative clause. The model predicted this effect, however, only on the determiner, since this is the point where the invalidity of the expectation becomes evident. The left panel of Figure 5.9 shows the difference between data and prediction.

5.3.1 Simulation

ACT-R contains a module for the articulatory loop (Baddeley, 2003; Baddeley & Hitch, 1974), which can be used as a temporary memory via subvocalization. Alternatively to an immediate time-out when the parser is occupied, a limited number of words could be held in articulatory memory before reading has to be interrupted. This has the advantage of a more flexible timing between parsing and eye movements, which is adaptive to current needs.

To test the idea of subvocalization, a simple model was implemented that could hold one item in the loop by subvocalization, thus delaying the necessity for a time-out regression by one word:

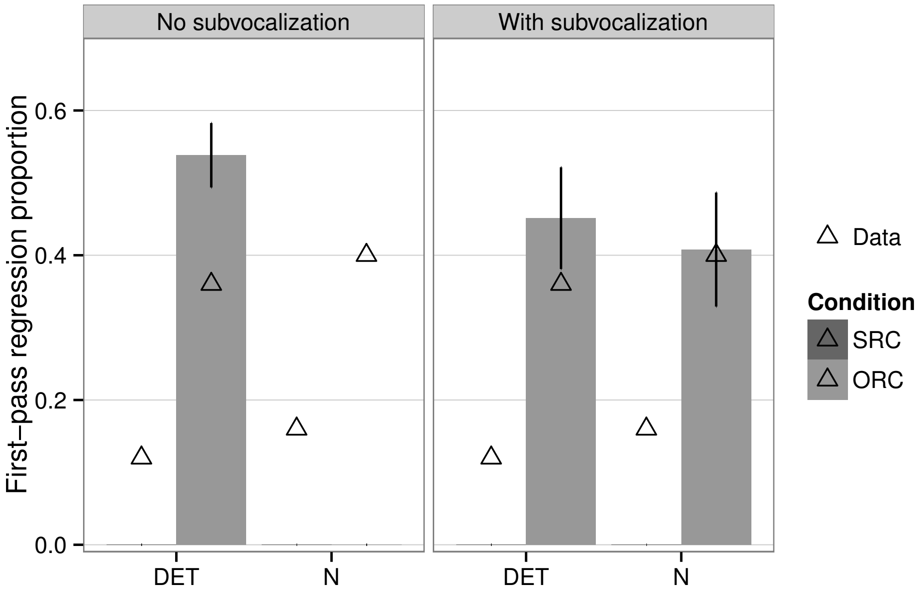

Figure 5.9: Detailed model predictions and Staub (2010a) data at embedded noun phrase. Left panel shows predictions of a model without subvocalization. Right panel shows predictions of a model using the articulatory loop for subvocalization.

Interface IV: Subvocalization If a word is ready for integration but the parser is occupied, integration can be delayed without interrupting the eye movements by holding the word in the articulatory loop until the next word is identified.

5.3.2 Results and Discussion

Running the subvocalization model on the Staub (2010a) sentences resulted in a distribution of the effect of expectation in the ORC across both the determiner and the noun with a remarkably good numerical replication of the data (see right panel of Figure 5.9).

This simulation shows the potential of a subvocalization model. However, at least two essential questions have to be answered for such an account: (i) When are words stored in phonological memory and when is a time-out necessary? (ii) What is the capacity of the articulatory loop in reading?

A plausible idea is that the amount held in articulatory memory is flexibly adapted to the combinatorial needs of the parser, thus collecting the elements within a predicted constituent before structurally combining them. Hence, both the clearing of the loop and most syntactic integration would be carried out when completing a constituent (Huey, 1908; Kleiman, 1975). This is consistent with observations of clause wrap-up effects, which show increased fixation duration not only at the end of sentences but also at comma-marked clause boundaries (Hirotani, Frazier, & Rayner, 2006; Rayner et al., 2000). In a model delaying parsing effort by subvocalization toward the completion of constituents, time-out regressions would appear mostly before clause boundaries. It would thus predict the observation that regressions are less likely to occur across a boundary than within clauses (Hirotani et al., 2006).

To summarize, a model of subvocalization as temporary storage within constituents would make the following predictions:

- Delayed effects of parsing difficulty (spill-over) depending on clause length.
- Wrap-up effects appear at phrasal boundaries due to clearing the loop.
- Time-out regressions are most likely at clause boundaries.

In order to test this model, it is necessary to run experiments of reading structurally challenging sentences while manipulating the availability of subvocalization, e.g., by articulatory suppression (Baddeley et al., 1981; Daneman & Newson, 1992; Eiter & Inhoff, 2010; Kleiman, 1975; Slowiaczek & Clifton Jr., 1980).

5.4 General Discussion

In this chapter three additional interfaces between parsing and eye movement control have been defined and tested. Including the Time Out interface introduced in Chapter 4, the framework now comprises four interfaces, which are summarized below:

I. Time Out: Short regressions compensate for slow syntactic integration (Chapter 4, simulation on Potsdam Sentence Corpus)

II. Reanalysis: Immediate attention shift to previous material when structure has to be revised (Section 5.1, simulation of Staub, 2010a data)

III. Underspecification: For structural relations with low utility, time-consuming attachments are aborted as soon as the next word is ready for integration, such that reading proceeds uninterrupted (Section 5.2, simulation of von der Malsburg & Vasishth, 2013, data)

IV. Subvocalization: A limited number of words can be stored in the articulatory loop in order to delay integration and to ensure uninterrupted reading (Section 5.3)

The first two interfaces, Time Out and Reanalysis, cause parser-triggered *interruptions* of the reading process, whereas Underspecification and Subvocalization provide mechanisms that abort or delay parsing to ensure *uninterrupted* reading. Time Out is a relatively straight-forward implementation of a well-defined mechanism. However, the other three interfaces are simplifying descriptions of more complex mechanisms. The Reanalysis interface leaves open the question where regressions are targeted and how the target position is determined. This point will be addressed in the Outlook of Chapter 6. Interfacte III predicts *when* attachment is aborted due to an interaction of parsing difficulty, word identification timing, and individual differences, but it lacks a definition of *which* relations are eligible candidates for potential underspecification. As discussed above, construal theory provides an orientation for this question, but a continuous experience- and context-based utility value for syntactic relations would be a more realistic model. Finally, Interface IV: *Subvocalization* is rather speculative and needs empirical testing.

Chapter 6

Summary and Conclusion

The primary goal of this thesis was to develop a model that integrates several aspects of reading that so far have been studied only separately. Using the domain-independent architecture of cognitive processing ACT-R, an established theory of cue-based retrieval was combined with a parsing mechanism and a model of eye movement control. The rationale behind this project is that behavioral measures of sentence processing such as fixation durations and regressions in reading are not direct indicators for processing difficulty but are the result of an interaction of multiple processing levels and strategies. In particular, a rationally bounded trade-off between reading speed and depth of comprehension can only be explained adequately if the interactions between the factors involved are sufficiently understood. Generating predictions according to interactions across multiple cognitive levels is only possible with explicit definitions of the respective mechanisms as well as their causal relationship with observable behavior. This thesis makes several contributions toward a model that provides these requirements.

6.1 Summary

Chapter 3 The model developed here is embedded into ACT-R (Anderson et al., 2004) and depends on the theory of cue-based retrieval as described in Lewis and Vasishth (2005). Chapter 3 therefore pursues a thorough assessment of the model's coverage of empirical data and points out discrepancies between verbally stated predictions used in the literature and the actual behavior of the mathematical implementation of the model. A comprehensive literature review of retrieval interference in reflexive-antecedent dependencies, number agreement, and non-agreement subject-verb dependencies is reported, and the predictions of cue-based retrieval theory are computationally evaluated with respect to published results. A novel finding from the review and modeling is that, contrary to claims in the literature, results on number agreement are not entirely compatible with cue-based retrieval theory. It is also shown that the cue-based retrieval account in its current form cannot explain several reported interference effects, such as (i) speed-ups observed in presence of a syntactically unlicensed distractor when the correct dependent is a full match to the retrieval cues and (ii) slow-downs when the correct dependent only partially matches the retrieval cues. It is demonstrated that these effects can be explained by the independently motivated theoretical constructs of *distractor prominence* and *cue confusion.*

By implementing a *prominence correction* and mapping prominence to distractor base-level activation, *distractor prominence* accounts for the influence of experimental design on the magnitude and the direction of the interference effect in target-match conditions. As an independently motivated principle, it is proposed that cues do not match features in a one-to-one fashion, but cues are *associated* with multiple features to different degrees as the result of usage-based learning. This predicts that in certain linguistic environments and possibly as the result of an adaptation to limited resources, associations between certain features and retrieval cues can be *underspecified*, i.e., cues can be *confused.*

The extended cue-based retrieval model is shown to provide a better explanation of published results than the classical retrieval account. Because the motivations for the two newly introduced principles are independent from a specific task, these principles are proposed as a general extension to the ACT-R theory of cue-based retrieval.

Chapter 4 The goal of Chapter 4 is to develop and test a framework for modeling the interaction of cue-based retrieval parsing and eye movement control. The essence of the framework is the implementation of a straight-forward interface which is inspired by E-Z Reader 10 (Reichle et al., 2009) and the Time Out hypothesis of Mitchell et al. (2008) describing the primary function of short-distance regressions. Using the parsing model of Lewis and Vasishth (2005) and the eye movement model EMMA (Salvucci, 2001) in ACT-R, the *Time Out* interface is implemented as the single possibility for the parser to intervene with oculomotor control, which otherwise proceeds autonomously. The Time Out function initiates a short regression when a word has been lexically identified but the parser is not ready to integrate that word because it is still occupied with the integration of previous input. The regression ensures that no new input is processed until the parser has caught up with processing.

As a preparatory step, Salvucci (2001)'s simulations with EMMA on the English Schilling Corpus (Schilling et al., 1998) are replicated and parameters are re-estimated for the ACT-R 6.0 environment. Then, the Time Out model is thoroughly tested on the German Potsdam Sentence Corpus (Kliegl et al., 2004), using pre-computed values for retrieval parsing (Lewis & Vasishth, 2005) and surprisal (Levy, 2008) as provided in Boston et al. (2011) and Boston et al. (2008), respectively.

The simulations show three major results: (i) The Time Out interface enables the model to predict regression-related measures such as rereading time and first-pass regression probability as a function of post-lexical processing that pattern with the corpus data. (ii) The ability to produce inter-word regressions also improves the predictions for fixation measures which are not directly related to regressions. (iii) The simulation results corroborate the findings of Boston et al. (2011) that retrieval and surprisal are complementary predictors of eye movement behavior, affecting both early and late measures.

In conclusion, the modeling supports Mitchell et al. (2008)'s assumption that the majority of short-distance regressions in sentence processing have the function to compensate for slow post-lexical processing, and it demonstrates that processes that occur *late* can nevertheless influence fixation measures that are assumed to reflect *early* processes. Most importantly, however, the study shows that the choice of ingredients to the model (ACT-R, cue-based parsing, and Time Out) results in an elementary framework suitable for further studying the eye-parser interaction.

Chapter 5 Finally, Chapter 5 complements Chapter 4 in that it establishes three additional eye-parser interfaces, which, including Time Out, are proposed as an elementary set to capture the interactions of cognitive processing and eye movement behavior in sentence processing. Interface II: *Reanalysis* triggers a regression whenever the parser, based on unexpected input, needs to reanalyze previously built structure. Interface III: *Underspecification*, in contrast to interfaces I and II, does not interrupt the reading process but rather ensures uninterrupted reading by abandoning time-consuming parsing decisions if they are not essential for the sentence interpretation (Frazier & Clifton, 1997). In these cases, autonomous saccade programming constrains the time available for an immediate attachment. In combination with the assumption that attachment time is affected by limited resources, this leads to the prediction that time-consuming attachment decisions are canceled mostly by readers with low working memory capacity. Finally, Interface IV: *Subvocalization* is proposed as a general means to decouple parsing further from the currently attended word by using the articulatory short-term storage (Baddeley, 2003).

Using the parameter values estimated in Chapter 4, the predictions of the proposed eye-parser interfaces are tested in simulations that use the Lewis and Vasishth (2005) parser in runtime rather than pre-computed values. A simulation of reading subject- and object-relative clauses shows that Interfaces I and II accurately predict the findings of Staub (2010b) on qualitatively different effects of memory and expectation, manifesting themselves as increased fixation durations and increased first-pass regression probability, respectively. Interface IV: *Subvocalization* improves the prediction of spill-over effects in the same simulation. The predictions of Interface III: *Underspecification* are tested in reading ambiguous adjunct attachments in Spanish (von der Malsburg & Vasishth, 2013). The simulation of participants with randomly assigned working memory capacity results in more regressions due to reanalysis for high-capacity readers and a greater ambiguity advantage for low-capacity readers in the ambiguous region. Both findings are consistent with the empirical results reported by von der Malsburg and Vasishth (2013).

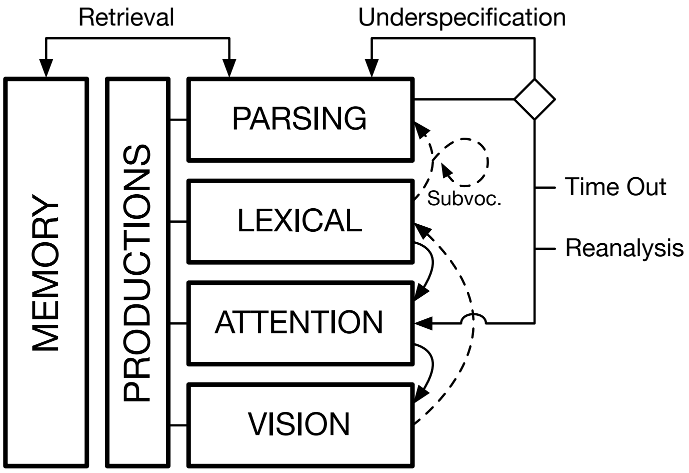

Figure 6.1: Structure of the extended ACT-R model showing interactions (solid lines) and information flow (dashed lines) between components during reading.

The results presented in Chapter 5 thus demonstrate that (i) the parameters estimated on a German corpus provide precise quantitative predictions for English relative clauses; (ii) different rates of underspecification for high- and low-capacity readers can emerge from a common trade-off mechanism between reading speed and processing effort rather than from different underlying strategies; and (iii) the four proposed interfaces prove suitable as an elementary set to capture important findings in the literature on sentence processing in terms of interactions between eye movement control, parsing, and individual resources.

Figure 6.1 provides an overview of the structure of the final model as implemented in ACT-R 6. There is a basic interaction loop between attention, vision, and lexical encoding, depicted by curved solid (direct interaction) and dashed (information flow) arrows: Attention directs the eye movements to a certain target. Visual information about the target is processed by the lexical module, which, in turn, allows the next attention target to be selected. Lexical information flows from the lexical module to the parsing module with the phonological loop used for short-term storage through subvocalization. The parsing module constantly interacts with memory by storing linguistic representations and retrieving those on the basis of heuristics (the retrieval cues) in order to complete dependencies. If parsing difficulty causes a processing backlog or failure, the parser has two options, depicted by the diamond: One option is to abort the current process (Interface III: *Underspecification*). This option is pursued whenever the current process seems expendable for the level of representational preciseness that is required for the current reading goal. The second option is to interfere with eye movement control by redirecting visual attention. In the case of a processing backlog, this would be achieved by a Time Out regression (Interface I). In the case of a parsing breakdown, a reanalysis (Interface II) is necessary.

6.2 Outlook

In this thesis, a basic framework is developed that integrates prominent findings in the literature with independently motivated principles. However, more thorough testing and refinement is necessary, which is beyond the scope of this work. The model in its current form incorporates a great range of interactions between mechanisms of sentence comprehension that can and should be assessed experimentally. Especially, the predictions of *cue confusion* and *distractor prominence*, of Interface III: *Underspecification*, and the interplay of *Time Out* and *Subvocalization* are particularly interesting for designing future experiments.

The restriction of the current framework to syntactic processing is obviously a simplification. It is undeniable that higher cognitive levels such as semantics and discourse context play an important role in sentence processing. A relevant cognitive model in this context is the work of Budiu and Anderson (2004), who modeled contextual effects on sentence processing in ACT-R using a compositional semantic representation of propositions. In principle, the EMMA/ACT-R model could be augmented in a similar way. The tree structure built by the Lewis and Vasishth (2005) parser encodes basic relations necessary to understand a proposition, which in principle makes it possible to derive semantics from the tree.

Furthermore, as EMMA is a domain-independent vision module, timing and location of single fixations is not as precise as in more sophisticated models like E-Z Reader or SWIFT. For example, in E-Z Reader, forward movement is more autonomous, such that a saccade program starts before the actual attention shift. In EMMA, saccade programming currently starts simultaneously with the attention shift, but with some changes in the code it should be possible to refine the timing of saccade planning and attention to better account for current evidence in reading research.

6.2.1 The *Why*, *Where*, and *How* of regressions

The presented model currently specifies *when* regressions occur, and it is assumed that time-out regressions only serve the purpose of slowing down the eyes (Mitchell et al., 2008). But there is still much to learn about regressions that serve a different function than a time-out and that target earlier points in the sentence than the last fixated word, for example, when reanalyzing part of the sentence (Booth & Weger, 2013; Frazier & Rayner, 1982; Inhoff & Weger, 2005; von der Malsburg & Vasishth, 2011, 2013; Meseguer et al., 2002; Mitchell et al., 2008; Weger & Inhoff, 2007).

The most straight-forward answer to *why* long-range regressions occur would be that they serve to revisit and reintegrate information. Evidence for this has been found, for example, by Booth and Weger (2013). However, another possibility is that attending to earlier words is just a way to reactivate internal representations which are associated with spatial positions, and thus are easier to retrieve when attending to these positions. This is claimed, for example, by the "deictic pointer hypothesis" (Ballard, Hayhoe, Pook, & Rao, 1997; Spivey, Richardson, & Fitneva, 2004, see also Ferreira, Apel, & Henderson, 2008; Kennedy, 1992; O'Regan, 1992).

It is also an open question *how* a particular location in a sentence can be found. The easiest way to do it would be to search backward word by word. However, as has been pointed out already by Kennedy and Murray (1987), readers are able to target positions distant from the current fixation quite precisely. But how is this possible? Do readers store the spatial positions of all words (e.g., Kennedy, 1992; Kennedy, Brooks, Flynn, & Prophet, 2003; Kennedy & Murray, 1987; Zechmeister, McKillip, Pasko, & Bespalec, 1975), or are the positions reconstructed from verbal memory, i.e., by mapping the verbal length of sentence to spatial distance (see, e.g., Inhoff & Weger, 2005; Rawson & Miyake, 2002; Weger & Inhoff, 2007)? The experiments of Inhoff and Weger (2005) and Weger and Inhoff (2007) studied the influence of both possibilities, which they termed "spatial coding" and "verbal reconstruction", respectively. The results implied a mixture of both strategies being applied in finding long-distance targets in a sentence: Initial long regressions where guided by spatial memory and were not very precise, landing in the proximity of the target. These regressions were followed by shorter regressions that directed the eyes onto the target word and were influenced by verbal memory, i.e., they were more precise when their target was a word that occurred early in the sentence. Inhoff and Weger (2005) hypothesized that "words that occurred early in a sentence may be the beneficiaries of primacy, as are words that occur early in a sequence of to-be-recalled words" (p. 460).

Regarding the storage of spatial information, it seems that not every word representation has an exact location stored with it. According to the spatial-indexing hypothesis of Kennedy and Murray (1987) and Kennedy (1992), stored locations ("spatiotopic values") could also be assigned to phrases or some cognitive processing that took place at a specific location. Similar accounts are the visuo-spatial sketchpad in the model of working memory by Baddeley (2003) or Ballard et al. (1997)'s model of pointers to spatial locations. A particularly useful model, which has been made a part of ACT-R, is the spatial indexing model of Pylyshyn (1989). It postulates so-called FINSTs ("fingers of instantiation"), which provide referents to visual features even if they are outside the visual field. According to Pylyshyn, four to five moving objects (including the currently attended one) can be indexed in this way simultaneously. The FINST model could be used in future simulations in ACT-R in combination with verbal information to generate predictions for regression paths from different weightings of spatial and linguistic factors.

6.2.2 Expectation-Based Serial Parsing in ACT-R

Probabilistic expectation-based models of parsing difficulty, such as surprisal (Hale, 2011; Levy, 2008), define a quantitative relation between corpus frequencies and an abstract notion of difficulty per word in a sentence. Ultimately, however, the mechanisms of expectation-based processing should be operationalized in an explicit parsing model. This will help to understand the relation between expectation and behavior such as, for example, the results of the simulations in Chapter 4 and evidence from Boston et al. (2011) that indicate that both early and late parsing processes are affected by structural expectation.

A possible translation of surprisal (Hale, 2011; Levy, 2008) in terms of ACT-R could be that rare combinations of parsing rules are executed more slowly than more frequent sequences. This would be the case under the assumption that more frequently used parsing rules remain more highly activated and thus fire faster. Such an approach would ground surprisal in procedural preferences trained by reading experience.

However, an alternative (or additional) approach involves an adaptive prediction mechanism that realizes surprisal as a mild form of garden-path. The following assumptions could be made for a predictive serial parser which makes predictions that are comparable to surprisal.

Assume a serial parser that has been trained on a treebank to predict upcoming structure at each point in a sentence. After integrating a word, It would now pre-build the most likely sentence structure in order to speed-up the integration of subsequent words. Assume further that the size of the pre-built structure depends on a trade-off between (A) the level of constraining information provided by the current context and (B) current memory load: (A) A context that highly constrains the possible continuations of a sentence allows the parser to increase the length of the predicted structure. However, (B) pre-building structure needs time, which is a limited resource per word shared with other parsing operations and memory processes, so that the necessity of other time-consuming processing would limit the amount of structure that can be built. When a prediction has been built and a new word is inconsistent with it, the predicted structure is discarded and new structure is created, similar to the recovery from a garden-path.

The predictions of a mechanism like this would be in line with findings in the literature. The primary prediction is the widely accepted observation that material inconsistent with the most probable structure leads to increased processing time (Hale, 2011; Levy, 2008). From the trade-off between (A) and (B) follows that stronger effects are predicted in material preceded by highly constraining context, and weaker effects are predicted in structures inducing high working memory load. This has been claimed by Gibson (2007) and Vasishth and Drenhaus (2011). Furthermore, as inconsistencies can be detected early by recognizing the input word category, expectation effects can occur earlier in the eye tracking record than working memory-related effects (Boston et al., 2011; Vasishth & Drenhaus, 2011). Finally, because predictions are realized as structural relations of predicted syntactic objects represented in working memory, discarded predictions remain available with the possibility to later interfere with the sentence interpretation (Staub, 2007).

6.3 Conclusion

This thesis makes the following major contributions: First, the predictions of the ACT-R-based theory of cue-based retrieval are thoroughly assessed with respect to interference in dependency processing and additional independently motivated principles are proposed, providing a more adequate model of cue-based retrieval.

Second, the definition of four elementary eye-parser interfaces and their implementation in a comprehensive model of eye movements in sentence processing provides a framework for further studies on the interaction of parsing and eye movement behavior.

Third, on the basis of the fist and second point, adaptive mechanisms of a trade-off between speed and effort are tested on two levels: On the level of memory access, the theory of *cue confusion* explains how associations between retrieval cues and features of memory items are adaptive to linguistic environments and individual differences, producing behavior that is unexplained under the classical cue-based retrieval theory. On the level of eye movements and syntactic processing, a cut-off mechanism is proposed which predicts empirical findings related to *good-enough processing* in reading such as the ambiguity advantage and individual differences in underspecification.

Fourth, a parsing module is developed (see Appendix A) and made publicly available as a general extension to ACT-R providing functions and special buffers for simulating sentence comprehension.

This thesis argues that, on all levels, the sentence comprehension mechanism seeks a balance between necessary processing effort and reading speed on the basis of experience, task demands, and resource limitations. Theories of sentence processing therefore need to be defined more explicitly, in particular with respect to linking assumptions between observable behavior and cognitive processes. The model developed here provides explicit mechanisms that cause complex and adaptive behavior. It thus constitutes a comprehensive model integrating multiple levels of sentence processing that hitherto have only been studied in isolation. The model is made publicly available as an expandable framework for future studies of the interactions between parsing, memory access, and eye movement control.

References

Acuña–Fariña, J. C., Meseguer, E., & Carreiras, M. (2014). Gender and number agreement in comprehension in Spanish. *Lingua*, *143*, 108–128.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–60.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ariel, M. (1990). *Accessing noun-phrase antecedents.* London, UK: Routledge.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, *31*, 137–162.

Baayen, R., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical data base on CD-ROM.* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baddeley, A. D. (1979). Working memory and reading. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing of visible language* (Vol. 13, p. 355-370). Springer US. doi: 10.1007/978-1-4684-0994-9_21

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839.

Baddeley, A. D., Eldridge, M., & Lewis, V. (1981). The role of subvocalisation in reading. *The Quarterly Journal of Experimental Psychology*, *33*(4), 439–454.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

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

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(04), 723–742.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1178–1198.

Berwick, R., & Weinberg, A. (1984). *The grammatical basis of linguistic performance.* Cambridge, MA: MIT Press.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1168–1178). Uppsala, Sweden: Association for Computational Linguistics.

Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, *8*(1), 57–99.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive psychology*, *23*(1), 45–93.

Booth, R. W., & Weger, U. W. (2013). The function of regressions in reading: Backward eye movements allow rereading. *Memory & Cognition*, *41*(1), 82–97.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349.

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *10*(2), 137–167.

Budiu, R., & Anderson, J. (2004). Interpretation-based processing: A unified theory of semantic sentence comprehension. *Cognitive Science*, *28*(1), 1–44.

Carminati, M. N. (2005). Processing reflexes of the feature hierarchy (person> number> gender) and implications for linguistic theory. *Lingua*, *115*(3), 259–285.

Carreiras, M., & Clifton, C. (1993). Relative clause interpretation preferences in spanish and english. *Language and Speech*, *36*(4), 353-372. doi: 10.1177/002383099303600401

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and topic* (pp. 25–56). New York: Academic Press.

Chen, Z., Grove, K., & Hale, J. T. (2012). Structural expectations in Chinese relative clause comprehension. In J. Choi, E. A. Hogue, J. Punske, D. Tat, J. Schertz, & A. Trueman (Eds.), *Proceedings of the 29th West Coast Conference on Formal Linguistics* (pp. 29–37). Somerville, MA: Cascadilla Press.

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

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use.* New York, NY: Praeger.

Clackson, K., Felser, C., & Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: Evidence from eye-movements during listening. *Journal of Memory and Language*, *65*, 128–144.

Clackson, K., & Heyer, V. (2014). Reflexive anaphor resolution in spoken language comprehension: Structural constraints and beyond. *Frontiers in Psychology*, *5*(904).

Clifton, C., Frazier, L., & Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivista di Linguistica*, *11*(1), 11–39.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. Van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–372). Oxford, UK: Elsevier.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowles, H. W., Walenski, M., & Kluender, R. (2007). Linguistic and cognitive prominence in anaphor resolution: Topic, contrastive focus and pronouns. *Topoi*, *26*, 3–18.

Cunnings, I., & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, *28*(1-2), 188–219.

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139.

Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, *25*(3), 315–353.

Daneman, M., & Newson, M. (1992). Assessing the importance of subvocalization during normal silent reading. *Reading and Writing*, *4*(1), 55–77.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and*

Language, *69*, 85–103.

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language*, *63*, 805–855.

Du Bois, J. W. (2003). Argument structure: Grammar in use. In J. W. Du Bois, L. E. Kumpf, & W. J. Ashby (Eds.), *Preferred argument structure: Grammar as architecture for function* (Vol. 14, pp. 11–60). Amsterdam, The Netherlands: John Benjamins.

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language*, *36*, 147–164.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531.

Eiter, B. M., & Inhoff, A. W. (2010). Visual word recognition during reading is followed by subvocal articulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 457.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L., Hare, M., & McRae, K. (2004). Cues, constraints, and competition in sentence processing. In M. Tomasello & D. I. Slobin (Eds.), *Beyond nature–nurtur: Essays in honor of Elizabeth Bates* (pp. 111–138). Mahwah, NJ: Lawrence Erlbaum Associates.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2015). *The determinants of retrieval interference in dependency resolution: Review and computational modeling.* (Manuscript submitted)

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, *5*, 452—474. doi: 10.1111/tops.12026

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

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Sciences*, *12*(11), 405–410.

Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15.

Francis, W., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Franck, J., Colonna, S., & Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in Psychology*, *6*(349).

Franck, J., Vigliocco, G., & Nicol, J. (2002). Attraction in sentence production: The role of syntactic structure. *Language and Cognitive Processes*, *17*(4), 371–404.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1139–1144). Amsterdam, Netherlands: Cognitive Science Society.

Frazier, L. (1987a). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *The psychology of reading* (Vol. 12, pp. 559–586). Hillsdale, NJ: Erlbaum.

Frazier, L. (1987b). Syntactic processing: Evidence from dutch. *Natural Language & Linguistic Theory*, *5*(4), 519–559.

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

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

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.

Fukumura, K., & van Gompel, R. P. G. (2011). The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, *26*(10), 1472–1504.

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*(1), 1–23.

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*, 699–717.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press.

Gibson, E. (2007). Locality and anti-locality effects in sentence comprehension. In *Workshop on processing head-final languages.*

Gibson, E., Desmet, T., Grodner, D., Watson, D., & Ko, K. (2005). Reading relative clauses in english. *Cognitive Linguistics*, *16*, 313–353.

Gibson, E., & Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, *28*(1-2), 125–155.

Givón, T. (1983). *Topic continuity in discourse.* Amsterdam, The Netherlands: John Benjamins.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, *29*, 261–291.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd meeting of the North American chapter of the Association for Computational Linguistics* (pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hale, J. T. (2011). What a rational parser would do. *Cognitive Science*, *35*(3), 399–443.

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, *54*(3), 425–443.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*(1), 3–27.

Huey, E. B. (1908). *The psychology and pedagogy of reading.* The Macmillan Company.

Inhoff, A., & Weger, U. W. (2005). Memory for word location during reading: Eye movements to previously read words are spatially selective but not precise. *Memory & Cognition*, *33*(3), 447-461.

Jäger, L., Benz, L., Roeser, J., Dillon, B., & Vasishth, S. (2015). Teasing apart retrieval and encoding interference in the processing of anaphors. *Frontiers in Psychology*, *6*(506).

Jäger, L., Chen, Z., Li, Q., Lin, C.-J. C., & Vasishth, S. (2015). The subject-relative advantage in Chinese: Evidence for expectation-based processing. *Journal of Memory and Language*, *79*, 97–120.

Jäger, L., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, *6*(617).

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149.

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: An ERP study. *Journal of Psycholinguistic Research*, *31*(2), 165–193.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *8*(1), 63–99.

Kemper, S., Crow, A., & Kemtes, K. (2004). Eye fixation patterns of high and low span young and older adults: Down the garden path and back again. *Psychology and Aging*, *19*, 157–170.

Kennedy, A. (1992). The spatial coding hypothesis. In *Eye movements and visual cognition* (pp. 379–396). Springer.

Kennedy, A., Brooks, R., Flynn, L.-A., & Prophet, C. (2003). The reader's spatial code. *The mind's eye: Cognitive and applied aspects of eye movement research*, 193–212.

Kennedy, A., & Murray, W. S. (1987). Spatial coordinates and reading: Comments on Monk (1985). *The Quarterly Journal of Experimental Psychology*, *39*(4), 649–656.

King, J., Andrews, C., & Wagers, M. (2012). Do reflexives always find a grammatical antecedent for themselves? In *25th annual CUNY conference on human sentence processing* (p. 67). New York, NY: The CUNY Graduate Center.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Kleiman, G. M. (1975). Speech recoding in reading. *Journal of Verbal Learning and Verbal Behavior*, *14*(4), 323–339.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1), 262–284.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*(6), 627–645.

Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, *5*(1252).

Kwon, N., Gordon, P. C., Lee, Y., Kluender, R., & Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of prenominal relative clauses in Korean. *Language*, *86*(3), 546–582.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

Lau, E. F., Rozanova, K., & Phillips, C. (2007). Syntactic prediction and lexical surface frequency effects in sentence processing. *University of Maryland Working Papers in Linguistics*, *16*, 163–200.

Lehtonen, M., Niska, H., Wande, E., Niemi, J., & Laine, M. (2006). Recognition of inflected words in a morphologically limited language: Frequency effects in monolinguals and bilinguals. *Journal of Psycholinguistic Research*, *35*(2), 121–146.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

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. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

Lin, C.-J. C., & Bever, T. G. (2006). Subject preference in the processing of relative clauses in Chinese. In D. Baumer, D. Montero, & M. Scanlon (Eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics* (pp. 254–260). Somerville, MA: Cascadilla Press.

Logačev, P., & Vasishth, S. (2014). *What is underspecification?* (Manuscript submitted)

Logačev, P., & Vasishth, S. (2015). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*. (In press) doi: doi:10.1111/cogs.12228

Lovett, M. C., Reder, L. M., & Lebiere, C. (1999). Modeling working memory in a unified architecture. *Models of working memory: Mechanisms of active maintenance and executive control*, 135–182.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676.

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*(2), 109–127.

von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545–1578.

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

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 155–200). San Diego, CA: Elsevier.

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

Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, *30*(4), 551–561.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research*, *24*, 469–488.

Mitchell, D. C., Shen, X., Green, M. J., & Hodgson, T. L. (2008). Accounting for regressive eye-movements in models of sentence processing: A reappraisal of the Selective Reanalysis hypothesis. *Journal of Memory and Language*, *59*(3), 266–293.

New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language*, *51*(4), 568–585.

Newell, A. (1973). *Production systems: Models of control structures* (Tech. Rep.). DTIC Document.

Newell, A. (1978). *Harpy, production systems and human cognition* (Tech. Rep.). Carnegie Mellon University.

Nicol, J., Forster, K. I., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, *36*, 569–587.

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, *18*(1), 5–19.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*, 601–626.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *46*(3), 461.

Parker, D., & Phillips, C. (2014). Selective priority for structure in memory retrieval. In *27th Annual CUNY Conference on Human Sentence Processing* (p. 100). Columbus, OH.

Patil, U., Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2015). A computational evaluation of sentence processing deficits in aphasia. *Cognitive Science*. doi: 10.1111/cogs.12250

Patil, U., Vasishth, S., & Kliegl, R. (2009). Compound effect of probabilistic disambiguation and memory retrievals on sentence processing: Evidence from an eye-tracking corpus. In A. Howes, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling.* Manchester, UK: University of Manchester.

Patil, U., Vasishth, S., & Lewis, R. L. (2012). *Retrieval interference in syntactic processing: The case of reflexive binding in English.* (Manuscript submitted)

Pearlmutter, N. J. (2000). Linear versus hierarchical agreement feature processing in comprehension. *Journal of Psycholinguistic Research*, *29*(1), 89–98.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*, 427–456.

Phillips, C., Wagers, M., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner (Ed.), *Experiments at the interfaces* (Vol. 37, pp. 147–180). Bingley, UK: Emerald Group Publishing Limited.

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97.

R Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.R-project.org/`

R Core Team. (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.R-project.org/`

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Rawson, K. A., & Miyake, A. (2002). Does relocating information in text depend on verbal or visuospatial abilities? An individual-differences analysis. *Psychonomic Bulletin & Review*, *9*(4), 801–806.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology*, *53A*(4), 1061–80.

Reichle, E. D. (2015). Computational models of reading: A primer. *Language and Linguistics Compass*, *9*(7), 271–284.

Reichle, E. D., Pollatsek, A., Fisher, D., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4–22.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using E-Z Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye-mind link. *Psychological Review*, *119*(1), 155.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher-level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21.

van Rij, J., van Rijn, H., & Hendriks, P. (2013). How WM load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, *5*(3), 564–580.

RStudio, & Inc. (2014). shiny: Web Application Framework for R [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=shiny` (R package version 0.10.2.1)

Salvucci, D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, *1*(4), 201–220.

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

Schilling, H. E. H., Rayner, & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, *26*(6), 1270–1281.

Schriefers, H., Friederici, A. D., & Kuhn, K. (1995). The processing of locally ambiguous relative clauses in German. *Journal of Memory and Language*, *34*(4), 499–520.

Severens, E., Jansma, B. M., & Hartsuiker, R. J. (2008). Morphophonological influences on the comprehension of subject–verb agreement: An ERP study. *Brain Research*, *1228*, 135–144.

Slowiaczek, M. L., & Clifton Jr., C. (1980). Subvocalization and reading for meaning. *Journal of Verbal Learning and Verbal Behavior*, *19*(5), 573 - 582. doi: http://dx.doi.org/10.1016/S0022-5371(80)90628-3

Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. *The interface of language, vision, and action: Eye movements and the visual world*, 161–189.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521–1543.

Staub, A. (2007). The return of the repressed: Abandoned parses facilitate syntactic reanalysis. *Journal of Memory and Language*, *57*(2), 299–323.

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, *60*(2), 308–327.

Staub, A. (2010a). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86.

Staub, A. (2010b). Response time distributional evidence for distinct varieties of number attraction. *Cognition*, *114*(3), 447–454.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652–654.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*, 421–457.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542–562.

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.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*.

Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. *Memory and Cognition*, *35*(5), 1107–1121.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90.

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.

Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology*, *6*(347).

Van Dyke, J. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(2), 407–430.

Van Dyke, J., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316.

Van Dyke, J., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166.

Van Dyke, J., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.

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

Vasishth, S., Bruessow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*, 685–712.

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, *8*(10), e77006. doi: 10.1371/journal.pone.0077006

Vasishth, S., & Drenhaus, H. (2011). Locality in German. *Dialogue & Discourse*, *2*(1), 59–82.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794.

Vigliocco, G., Butterworth, B., & Semenza, C. (1995). Constructing subject-verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, *34*, 186–215.

Vitu, F., & McConkie, G. W. (2000). Regressive saccades and word perception in adult reading. In A. Kennedy, D. Heller, J. Pynte, & R. Radach (Eds.), *Reading as a perceptual process* (pp. 301–326). Oxford, UK: Elsevier.

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*, 206–237.

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, *14*(4), 770–775.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *104*(4), 442–452.

Weger, U. W., & Inhoff, A. (2007). Long-range regressions to previously read words are guided by spatial and verbal memory. *Memory & Cognition*, *35*(6), 1293–1306.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, *108*(1), 40–55.

Zechmeister, E. B., McKillip, J., Pasko, S., & Bespalec, D. (1975). Visual memory for place on the page. *The Journal of General Psychology*, *92*(1), 43–52.

Appendices

Appendix A

Parsing Module

The Lewis and Vasishth (2005) model has been implemented in ACT-R 5.0 and was for this thesis ported to the newer ACT-R 6.0. Besides some adaptation of syntax and naming, this involved the reduction of a range of model-specific customizations of general ACT-R code that provided special buffer behavior and circumvented the problem of chunk duplication that is described below. Instead of re-implementing these little 'hacks' in ACT-R 6.0, a proper ACT-R module was developed that provides special functions and buffers that behave in the way necessary for modeling sentence comprehension. The major functions of the module are explained below.

First, it introduces parallelism between parsing, lexical processing, and other cognition. In ACT-R, the retrieval buffer is a bottleneck that all memory retrieval operations use. It is, however, plausible that higher cognition, e.g., on the discourse level, or low-level processes such as word recognition proceed in parallel to the rapid retrieval operations of syntactic parsing. In order to provide this function, the parsing module contains two special retrieval buffers: `GRAMMATICAL` for retrieval of syntactic objects and `LEXICAL` for retrieval of lexical entries. Both buffers can carry out retrievals in parallel to each other and to the original `RETRIEVAL` buffer. Both new buffers have their own parameters for the latency factor F and the latency exponent f of Equation 2.5, so retrieval speed can be defined for structural and lexical retrieval separately from other cognition.

Both buffers are also special in the way that they avoid unwanted chunk duplication. In ACT-R, when a chunk (a memory item) is retrieved into a buffer, manipulated, and released into memory, a copy is created, leaving the original and the manipulated chunk in declarative memory. This seems a rational mechanism for a memory system, since we usually do not actively alter our memories. The mechanism of activation decay over time ensures that newly created chunk versions are more likely to be retrieved later than older versions. Chunks that are retrieved and released unaltered are merged with the originals. However, in the Lewis and Vasishth

(2005) model of retrieval parsing, syntactic objects that have just been created are retrieved and manipulated in order to build dependencies between these objects. Hence, despite decay, often the wrong duplicate of a chunk would be retried, resulting in parsing failure. It is therefore necessary to update chunks without creating duplicates. In the original Lewis and Vasishth (2005) model, this was done by altering ACT-R code. In the new parsing module, the standard retrieval buffer works in the usual ACT-R fashion. But the grammatical and lexical buffers have the option to *force-merge* the chunks released from the buffers with their originals.

Furthermore, the parsing module provides functions for maintaining and monitoring the parsing status. The code and a documentation can be found at: `https://github.com/felixengelmann/ACT-R-Parsing-Module`.

Appendix B

Software

An objective of this thesis project is to provide a framework with implemented assumptions that can be tested and falsified, or extended by other researchers in order to advance the understanding of the interaction of various levels in sentence comprehension. Therefore, the developed code is provided for further use and advancement in the following locations:

- The extended model of cue-based retrieval of Chapter 3 including *distractor prominence*, *cue confusion*, and *cue-weighting* can be used online for simulations at:
  `https://engelmann.shinyapps.io/inter-act`
- The parsing module described in Appendix A is ready to use with ACT-R 6 and available at:
  `https://github.com/felixengelmann/ACT-R-Parsing-Module`
- The ACT-R 6 framework providing the re-implemented parsing model of Lewis and Vasishth (2005), an adjusted version of EMMA, the newly developed parsing module, and implemented eye-parser Interfaces I, II, and III is available at:
  `https://github.com/felixengelmann/act-r-sentence-parser-em`

Erklärung der Urheberschaft

Hiermit erkläre ich an Eides statt, dass bei der Abfassung der vorliegenden Arbeit alle Regelungen guter wissenschaftlicher Standards eingehalten wurden. Weiter erkläre ich, dass die vorliegende Arbeit selbständig verfasst wurde und über die Beiträge meiner Koautoren hinaus, welche in der beiliegenden *Erklärung über die Beiträge zu Gemeinschaftsarbeiten* spezifiziert sind, keine Hilfe Dritter in Anspruch genommen wurde.

Felix Engelmann

Potsdam, November 2015