TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 5 (2013) 452–474

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12026

A Framework for Modeling the Interaction of Syntactic Processing and Eye Movement Control

Felix Engelmann, Shravan Vasishth, Ralf Engbert, Reinhold Kliegl

Department of Linguistics, University of Potsdam

Received 15 February 2012; received in revised form 12 December 2012; accepted 11 January 2013

Abstract

We explore the interaction between oculomotor control and language comprehension on the sentence level using two well-tested computational accounts of parsing difficulty. Previous work (Boston, Hale, Vasishth, & Kliegl, 2011) has shown that surprisal (Hale, 2001; Levy, 2008) and cue-based memory retrieval (Lewis & Vasishth, 2005) are significant and complementary predictors of reading time in an eyetracking corpus. It remains an open question how the sentence processor interacts with oculomotor control. Using a simple linking hypothesis proposed in Reichle, Warren, and McConnell (2009), we integrated both measures with the eye movement model EMMA (Salvucci, 2001) inside the cognitive architecture ACT-R (Anderson et al., 2004). We built a reading model that could initiate short "Time Out regressions" (Mitchell, Shen, Green, & Hodgson, 2008) that compensate for slow postlexical processing. This simple interaction enabled the model to predict the re-reading of words based on parsing difficulty. The model was evaluated in different configurations on the prediction of frequency effects on the Potsdam Sentence Corpus. The extension of EMMA with postlexical processing improved its predictions and reproduced re-reading rates and durations with a reasonable fit to the data. This demonstration, based on simple and independently motivated assumptions, serves as a foundational step toward a precise investigation of the interaction between high-level language processing and eye movement control.

Keywords: Sentence comprehension; Eye movements; Reading; Parsing difficulty; Working memory; Surprisal; Computational modeling

1. Introduction

In language comprehension research, most of the evidence about the cognitive processes involved comes from the study of eye movements in reading. As the reader's eyes

Correspondence should be sent to Felix Engelmann, Department of Linguistics, University of Potsdam, Haus 14, Karl-Liebknecht Str. 24–25, Golm D-14476, Germany. E-mail: felix.engelmann@uni-potsdam.de

move through a sentence, the sequence of fixations and their durations reflects the reader's allocation of attention and the processing effort necessary to combine the words incrementally into a coherent structure. The specific linking between fixation patterns and the underlying cognitive processes is, however, not trivial: Fixations are determined not only by immediate low-level processes like word recognition but also by more complex operations such as structural parsing decisions, contextual integration, and non-linguistic oculomotor constraints. In recent years, a number of computational models have emerged that help understanding the reading process in detail (e.g., Bicknell & Levy, 2010; Engbert, Longtin, & Kliegl, 2002; Engbert, Nuthmann, Richter, & Kliegl, 2005; Legge, Hooven, Klitz, Mansfield, & Tjan, 2002; Nilsson & Nivre, 2010; Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Pollatsek, & Rayner, 2006; Reilly & Radach, 2006). The two most developed models of this kind are E-Z Reader (Reichle et al., 2006) and SWIFT (Engbert et al., 2005). These generate predictions based on lexical variables like word frequency, word length, and cloze predictability. Although they differ fundamentally in their core assumptions about the nature of the reading process (E-Z Reader shifts attention serially, while SWIFT allows for parallel word processing guided by an attentional gradient), both models make very accurate predictions about when and where the eyes move. However, since these models rely on word-level information, their predictions are limited to rather simple sentences that do not induce severe interruptions of the reading process.

Postlexical processes like structural and semantic integration operate on a higher level and can only be uncovered by studying more complex sentences that contain long-range dependencies, ambiguities, or contextual inconsistencies. Challenging the sentence processor in this way reveals memory operations, structural and semantic predictions, and repair processes. In particular, there has been an abiding interest in identifying spatio-temporal distributions of short- and long-range regressions (backward saccades) in psycholinguistic literature (Frazier & Rayner, 1982; Meseguer, Carreiras, & Clifton, 2002; Mitchell et al., 2008; Van Dyke & Lewis, 2003; von der Malsburg & Vasishth, 2011, 2012; Weger & Inhoff, 2007). In most established eye movement models, however, interword regressions are caused either by incomplete lexical processing (e.g., SWIFT) or due to motor error (e.g., older versions of E-Z Reader). An exception is the model of Bicknell and Levy (2010), which explains regressions as the result of a rational strategy guided by Bayesian inference on the sentence level. The postlexical level of sentence processing has been captured by a range of computational models (e.g., Binder, Duffy, & Rayner, 2001; Budiu & Anderson, 2004; Elman, Hare, & McRae, 2004; Hale, 2011; Just & Carpenter, 1992; Konieczny & Döring, 2003; Lewis & Vasishth, 2005; MacDonald & Christiansen, 2002; Spivey & Tanenhaus, 1998; Vasishth, Bruessow, Lewis, & Drenhaus, 2008). These models predict word-by-word difficulty, which can be correlated with aggregated eye-tracking measures but abstracts away from individual fixations.

To understand how postlexical difficulty and eye movements interact, it is necessary to combine both classes of computational models and investigate the link between high-level language processes and oculomotor control. In a recent approach, Reichle et al. (2009) introduced a postlexical 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 to buy time for the integration process to finish. Although Reichle and colleagues did not integrate a computational account of postlexical 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, as follows: (a) The syntactic retrieval 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; (b) 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, 2010; Vasishth & Drenhaus, 2011), and corpus studies show that surprisal and retrieval are independent predictors of processing difficulty (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 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 eyetracking 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 article is to demonstrate the feasibility of integrating a computational account of postlexical 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. 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 eyetracking 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, Grabner, Rolfs, & Engbert, 2004) because measures of parsing difficulty are readily available for the latter. Section 2 will introduce EMMA in detail. In Section 3, 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. Finally, Section 5 presents six simulations on the German Potsdam Sentence Corpus that assess a range of model configurations that integrate EMMA with surprisal and retrieval.

2. The EMMA/ACT-R reading model

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, that is, 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:

$$T_{\text{enc}} = K(-\log f_i)e^{k\epsilon_i} \tag{1}$$

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 eyetracking measures.

Table 1
Frequency classes used in the analyses of the Schilling Corpus (SC) and Potsdam Sentence Corpus (PSC)

		SC		PSC	
Class	Freq. in 1M	Words	*M*	Words	*M*
1	1–10	77	3	186	3
2	11–100	87	50	173	41
3	101–1,000	71	333	200	335
4	1,001–10,000	92	5,067	207	5,020
5	> 10,000	112	41,976	84	2,399

3. Replication of Salvucci (2001)

3.1. 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 eyetracking measures for five logarithmic frequency classes (see Table 1). The frequency values available in the SC were obtained from Francis and Kucera (1982). 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 interword regressions (64%) were excluded from the analysis.

3.2. 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 to be fully integrated in ACT-R 6.0. All parameters except for those shown in Table 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.

3.3. 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

Table 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 five frequency classes for each eyetracking 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.

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 interword regressions. Then, we divided the corpus words into five logarithmic frequency classes (see Table 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 *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 in the Appendix. 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.

3.4. Results

We re-estimated the encoding factor K and the saccade preparation time T_{prep} to compensate for the changes in the ACT-R environment. See Table 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. 1 shows the predictions of the model (dashed lines) for six fixation measures as a function of frequency class. Besides the corpus data (gray 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.

3.5. 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. The extended EMMA/ACT-R model

To augment the EMMA/ACT-R model with postlexical 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

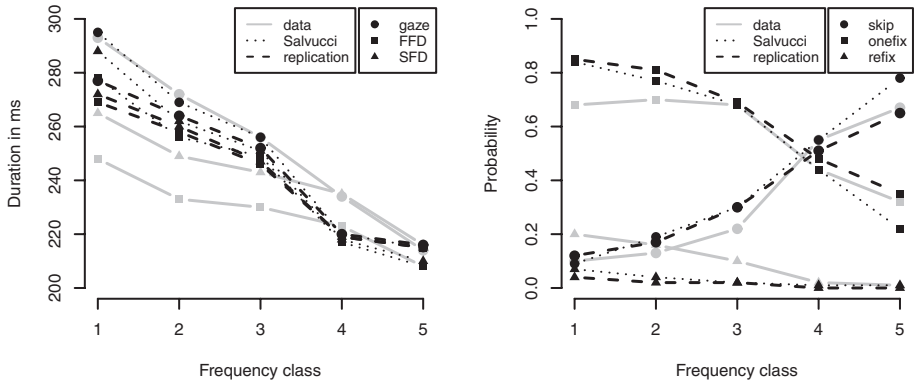

Fig. 1. Replication of Salvucci (2001) on the Schilling Corpus. Effects of word frequency on gaze, first, and single fixation duration, and on the rate of skipping a word, fixating it once and fixating it more than once. Gray solid lines represent experimental data, black dotted lines show Salvucci's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

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. Reichle and colleagues demonstrated the applicability of their model by re-configuring the model parameters for three cases of parsing difficulty: clause wrap-up, semantic violations, and garden paths.

Our goal is to evaluate a model that works in a similar way but uses a computational implementation of sentence comprehension to generate its predictions. 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, retrieval 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 possibility that retrieval fails completely (rapid integration failure) is not included in the model for now. Similar to E-Z Reader 10, when identification of word w_{n+1} finishes before the complete integration of word w_n, our model initiates a regression back toward the previous word. Once word integration is complete, the model continues with normal reading. This type of regression has been proposed by Mitchell et al. (2008). They called them "Time Out regressions" because their assumed function is to provide additional time for the sentence processor before taking up new input.

The above described 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 is 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.

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

4.1. Retrieval

In sentence processing, to create structural dependencies (e.g., between verbs and their arguments), items in memory have to be accessed; the success and duration of these access events are modulated by, inter alia, the distance between the dependents and the amount of interference from other items (Bartek, Lewis, Vasishth, & Smith, 2011; Gibson, 2000; Grodner & Gibson, 2005; Just & Carpenter, 1980, 1992). Lewis and Vasishth (2005) developed a computational model of parsing difficulty that adopts ACT-R's memory principles of fluctuating activation, decay over time, and similarity-based interference. The model was implemented in ACT-R as a fully specified left-corner parser that incrementally builds a structural representation, following X-bar rules. The constituents of the tree structure are stored as *chunks* in ACT-R's declarative memory (related to each other by features like specifier, comp, and head). To integrate an input word into the current structure, the parser carries out the following steps: (a) access the corresponding

lexical entry in the lexicon in declarative memory; (b) based on syntactic expectations, specify the features of a matching constituent and initiate a retrieval; and (c) create a new syntactic constituent and attach it to the one retrieved. Using the model's predictions of parsing duration, Lewis and Vasishth (2005) explained effects of distance and structural interference in sentence processing in terms of independently motivated principles of working memory access. The retrieval model has found further applications in accounts of anti-locality (Vasishth & Lewis, 2006), negative polarity constructions (Vasishth et al., 2008), reflexive binding (U. Patil, S. Vasishth, & R. Lewis, unpublished manuscript), and impaired sentence comprehension in aphasia (U. Patil, S. Hanne, S. Vasishth, F. Burchert, & R. De Bleser, unpublished manuscript).

To summarize, 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 to investigate the interaction between parsing and eye movements in detail. We connect this parsing model to EMMA via Time Out regressions.

4.2. Surprisal

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, for example, 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 postlexical processing. Given the conceptual distinctness of surprisal and retrieval together with experimental evidence locating expectation effects earlier than memory effects (Staub, 2010; 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" postlexical 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.

5. Simulations 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 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 postlexical 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 2 for an overview).

5.1. Data

5.1.1. Potsdam Sentence Corpus

The Potsdam Sentence Corpus contains eyetracking data from 144 simple German sentences (1,138 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 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.

5.1.2. 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 precalculated 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, that is, 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*.

5.1.3. 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[3] from Levy (2008).

5.2. 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 *integrateword* 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.

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 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. 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 2:

$$T_{\text{enc}} = (K[-\log f] + Ps)e^{k\epsilon} \tag{2}$$

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

5.3. Results

Simulation results are summarized in Table 2. Each model was evaluated on the prediction of frequency effects similar to the evaluation of the previous simulation (see Table 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), re-reading 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 re-reading a word after leaving it to the right (reread). Note that first-pass regression 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, all models were fit and evaluated on the full data set 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 1,000 ms and words with gaze and total fixation durations greater than 1,500 ms from empirical data set. 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.

5.3.1. PSC versus 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 2 shows the model performance when trials containing interword 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 ($R_{\text{early}} = 0.86$). 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 RMSDs are not directly comparable between corpora. RMSDs for the PSC are generally a bit lower because the standard deviations used for normalization are higher than in the Schilling Corpus.

5.3.2. Influence of parsing difficulty

In Table 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. To provide a baseline for the EMMA+ models, Simulation 2 was analyzed again on the complete data set, 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 of 85,000 (850 analyzed corpus words times 100 simulations) were re-read.

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. 2 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 surprisal and retrieval effects, we did a linear regression analysis for selected eyetracking measures using the predictors log frequency, length, log retrieval, and surprisal. This was done to see which of the six EMMA models produce variance that is explainable by surprisal and retrieval values. 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

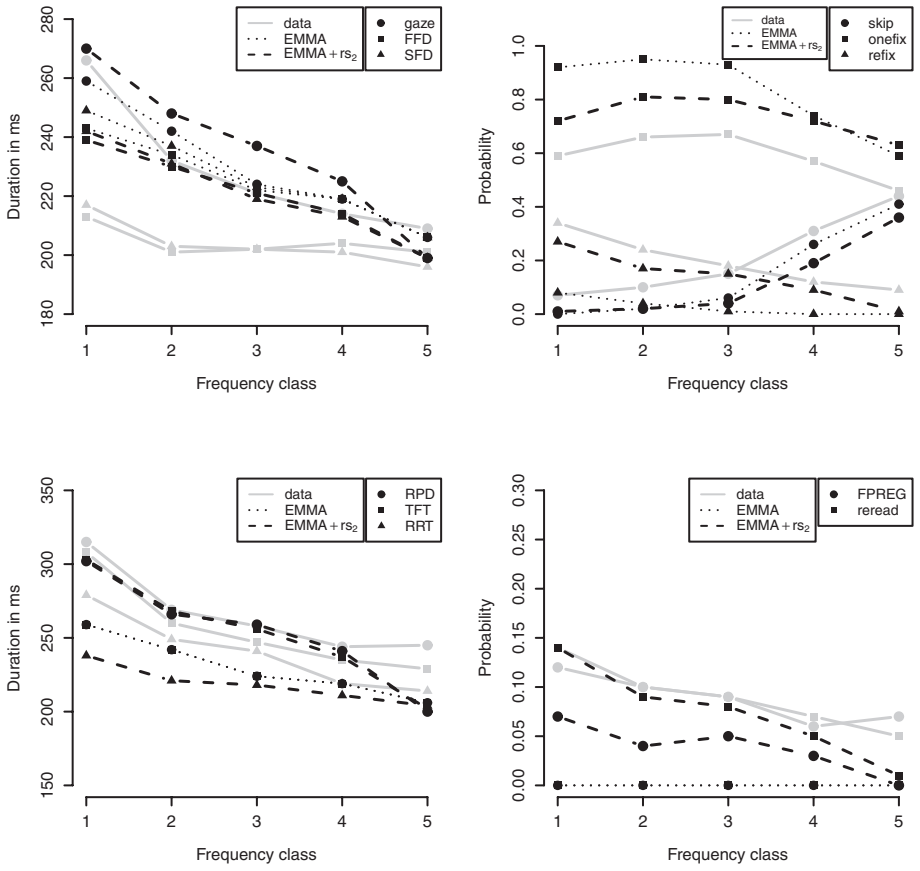

Fig. 2. Predictions of Model 2 (EMMA, dotted lines) versus Model 7 (EMMA+rs_2, dashed lines) versus experimental data (gray solid lines) for the Potsdam Sentence Corpus. The figure shows means of early (first row) and late measures (second row) as a function of frequency class. Each row shows reading time durations on the left and probabilities on the right side.

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. Overall, the regression analysis confirmed these expectations. More details about the analysis can be found in the Appendix.

5.4. Discussion

The results show that the extension with surprisal and retrieval information considerably improves EMMA's predictions for fixation measures. The interaction of postlexical 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.

6. General discussion

The primary goal of the current work 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 postlexical 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 re-reading time. The simulation results also 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, Vasishth, & Lewis, unpublished manuscript; Patil, Hanne, et al., unpublished manuscript; Vasishth & Lewis, 2006; Vasishth et al., 2008) can be further investigated using the published parsing rules seamlessly with the eye movement control model.

6.1. 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 postlexical 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. 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 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. Hence, the case of reanalysis is a good candidate for the application of what Reichle et al. (1998) call *rapid integration failure*, which cancels postlexical processing and initiates a regression. This would predict early effects and first-pass regressions in the disambiguating region in a garden path.

6.2. Future prospects

The restriction of the current framework to syntactic processing is obviously a simplification. It is undeniable that higher cognitive levels like semantics and 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. For the moment, however, we concentrate on syntactic effects. Our next step (this is work in progress) 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 future studies, the actual parsing architecture of Lewis and Vasishth (2005) will be used in runtime. As exemplified above, simulating explicit parsing processes at runtime enables the modeling of rapid integration failure in, for example, garden paths. Furthermore, it makes it possible to use linguistic information to define saccade targets. Short one-word regressions like the time-outs modeled here are very frequent and explain some of the variance in eye movement data. However, more complex regression patterns triggered by reanalysis have also been found (e.g., Frazier & Rayner, 1982; von der Malsburg & Vasishth, 2011, 2012; Meseguer et al., 2002). Readers often make long-range regressions to find the ambiguous region where the wrong attachment decision was made. Important questions regarding the eye-parser connection are to what degree are these long-range regressions guided by linguistic information and what is their exact function (e.g., Booth & Weger, 2013; Inhoff & Weger, 2005; Mitchell et al., 2008; Weger & Inhoff, 2007). In combination with the explicit ACT-R parsing model, EMMA can be used for studying these questions. Ultimately, expectation should also be modeled as a runtime process instead of being pre-calculated like surprisal. This will help to understand the nature of expectation-related effects. A possible translation of surprisal in terms of an ACT-R parser would be that rare combinations of parsing rules are executed slower

than more common sequences. Such an approach would ground surprisal in procedural preferences trained by reading experience.

To conclude, 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 eyetracking corpus not only statistically but also in an explicit computational model of eye movement control. With the presented framework, we plan to examine the individual contributions of surprisal and retrieval to the behavior at certain points of difficulty and the factors that guide long-range regressions.

Acknowledgments

We thank Dario Salvucci for providing us with information on his model and the original code. We also thank Hedderik van Rijn, Dan Bothell, and Mike Byrne for helpful clarifications on the re-implementation of EMMA, and Eric Reichle for releasing the raw Schilling Corpus data. This study was partly funded by the Deutsche Forschungsgemeinschaft.

Notes

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.
3. The parser is publicly available at http://nlp.stanford.edu/ ∼ rog/prefixparser.tgz
4. We concluded this by a recalculation of their values and personal communication with Dario Salvucci.

References

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), 1036–1060.

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

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

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.

Binder, K. S., Duffy, S., & Rayner, K. (2001). The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language*, *44*(2), 297–324.

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.

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

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

Elman, J. L., Hare, M., & McRae, K. (2004). Cues, constraints, and competition in sentence processing. In M. Tomasello and D. Slobin (Eds.), *Beyond Nature-Nurtur: Essays in honor of Elizabeth Bates*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42*(5), 621–636.

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.

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

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

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.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image language brain: Papers from the first mind articulation project symposium* (pp. 95–126). Cambridge, MA: MIT Press.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science: A Multidisciplinary Journal*, *29*(2), 261–290.

Hale, J. T. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (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.

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.

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.

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.

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In P. P. Slezak (Ed.), *Proceedings of the ICCS/ASCS joint International Conference on Cognitive Science* (pp. 330–335). Sydney, Australia: University of New South Wales.

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, S. J., & Tjan, B. S. (2002). Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Research*, *42*(18), 2219–2234.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

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

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.

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. (2012). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, in press, available online. doi:10.1080/01690965.2012.728232

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.

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.

Nilsson, M., & Nivre, J. (2010). Towards a data-driven model of eye movement control in reading. In J. T. Hale (Ed.), *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2010* (pp. 63–71). Uppsala, Sweden: Association for Computational Linguistics.

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, and R. P. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling*. Manchester, UK: University of Manchester.

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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), 1061–1080.

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., 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.

Reilly, R., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, *7*(1), 34–55.

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

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

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

Van Dyke, J. A., & 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*(3), 285–316.

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.

Vasishth, S., Bruessow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science: A Multidisciplinary Journal*, *32*(4), 685–712.

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.

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.

Appendix A: 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 squaringit.[4] 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:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{data_k - \mathrm{model}_k}{SD_k}\right)^2} \tag{3}$$

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

Appendix B: Linear regression analysis

To assess the contributions of surprisal and retrieval in the model, we performed a linear regression analysis. 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. 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). The models contained the predictors log frequency, length, log retrieval, and surprisal. See Equation 4 for an example.

$$\mathrm{FFD}_i = \beta_0 + \beta_1 log(\mathrm{freq}_i) + \beta_2 len_i + \beta_3 s_i + \beta_4 \log(r_i) + \epsilon_i \tag{4}$$

For each predictor, β is the coefficient to be estimated. Each of the predictors was additionally centered around zero. Fig. B1 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:

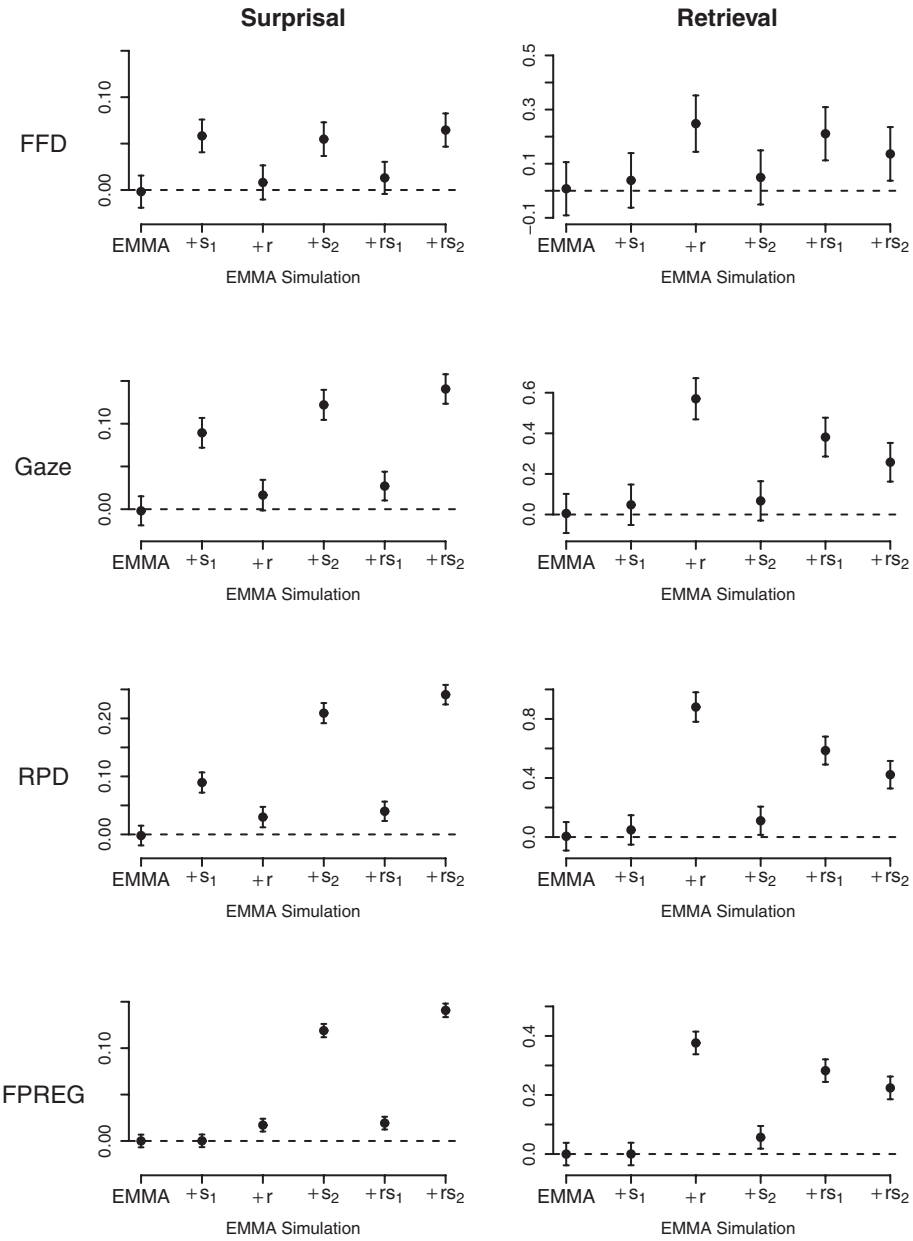

Fig. B1. Coefficients and 95% confidence intervals for predictors surprisal and retrieval estimated by linear regression. Predictors were log frequency, length, log retrieval, and surprisal. Coefficients are plotted along the *y*-axis for surprisal on the left side and retrieval on the right side. Regressions 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 $\alpha = 0.05$.

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 $r = .15$. 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 B1 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).

Table B1
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

Notes: 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).