Reinhold Kliegl and Ralf Engbert

Evaluating a Computational Model of Eye-Movement Control in Reading

1 Some basic facts about eye movements during reading

Reading is an activity we all engage in on a daily basis. It requires the coordination of perceptual and oculomotor processes as well as the integration of this new information provided by perception and eye movements with the available knowledge and expectations about what is being read. This chapter is about how we evaluate a model about an important component of this complex process, that is a model about how we move our eyes across the words of a sentence. We present this model as a prototype of an integrated theoretical, computational, and data-analytic approach for the interface between experimental psychology, cognitive (neuro-)science and computational neuroscience. With examples from ongoing research we illustrate how the model can be evaluated against a set of criteria for strong model tests, comprising goodness of fit, strictness of model, reliability of data, and unexpected predictions (Roberts and Pashler, 2000). Before we turn to these model tests we introduce some basic facts about eye movements during reading and (a subset of) the theoretical principles that guided the construction of our model.

When we record or directly observe a reader's eye movements, it is immediately apparent that the eyes do not move continuously across the words of a sentence.[1] Counter to what introspection suggests that we do most of the time; we notice a strict alternation of quick movements, called saccades (lasting about 20 to 30 ms), and periods of relative rest, called fixations (with mean durations ranging from 150 to 350 ms). Visual input about what we read occurs only during fixations; we are effectively blind while the eyes are in flight and indeed most of the time we are not aware of saccades. Thus, what we experience as a continuous movement across the words of a sentence is largely a construction of the mind.

1 The results reported in this chapter are based on binocular measurements at 250 Hz or 500 Hz from 273 readers who read 144 isolated sentences (i.e., the Potsdam Sentence Corpus; Kliegl, Grabner, Rolfs, and Engbert, 2004; Kliegl, Nuthmann, and Engbert, 2006). We map fixation positions to a specific letter within a word. The fixation positions differ slightly between the two eyes. Our results are based on measurements of the right eye.

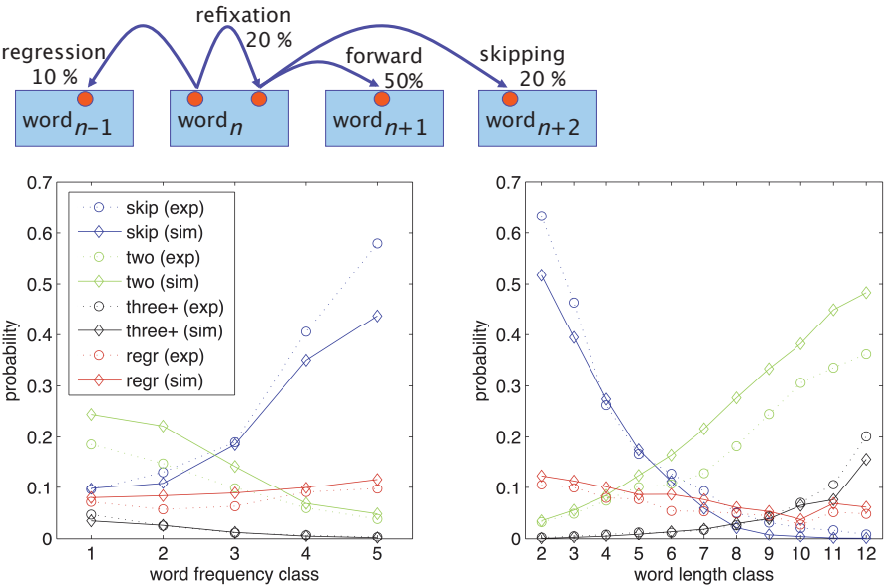

Figure 1: (top) Illustration of four types of eye movements and their marginal probabilities during normal reading of German prose. Experimental (exp) and simulated (sim) values for skipping (skip), regression (regr), and fixation probabilities (2 or >=3) conditional on word length (bottom left) and on log10 word frequency per million words (bottom right) (modified from Engbert et al., 2005).

The discrete nature of fixation-saccade cycles suggests a simple taxonomy of eye movements relative to the words of the sentence. As illustrated in Figure 1 (top panel), we distinguish four types: roughly 50% of the saccades carry the eyes from one word to the next word, about 20% of them shift the position within the currently fixated word, about 20% skip the next word, and about 10% of the time we move back to an earlier word. These statistics greatly depend on word properties, most notably on the lengths of the words. As shown in Figure 1 (bottom left), skipping probability decreases strongly from roughly 60% for two-letter words to less than 1% for 12-letter words; conversely the probability of refixations increases from close to 0% to around 40%. The same statistics can also be plotted over the frequency with which words are observed. As shown in Figure 1 (bottom right), skipping and refixation probability decreases and increases with log-frequency of observing a word in texts comprising one million words.

Fixation probabilities yield one key set of dependent measures. The second set of measures relates to different types of fixation durations. Here we distinguish, for example, between durations of fixations when they are the only fixa-

tion on a word (i.e., a single fixation duration), the first of two fixations, or the second of two fixations. Usually, we also sum all the fixations on a word to a measure of total reading time. Again, all these measures exhibit systematic relations to the lengths and frequencies of words on which they are observed: The longer or the less frequent a word, the longer the fixation duration or total fixation time.

The length and type frequency of the fixated word are but two highly correlated properties of a large number of variables that have been shown to influence fixation probabilities and durations (e.g., Rayner, 1998, for a review). Other variables are, for example, the predictability of a word given the prior words of the sentence. This measure is usually obtained in independent studies in which subjects have to guess the words of a sentence in an incremental order. Other variables reflect how similar the word is to other words of the language, measured by how many words can be derived from a word if one allows to exchange one letter (i.e., an edit distance of 1). Another example is the informativeness of the beginning of a word for its identification. For example, given a long word with "xy..." as initial letters, there are not many alternatives to "xylophone". Moreover and critically, properties of words n-1 or n+1 have been shown to influence fixation durations on word n (e.g., Kliegl et al., 2006). We will describe a few of these and a few additional effects in the context of introducing some of the theoretical principles guiding the evaluation of our computational model. In summary, despite the one-dimensional space during reading a single sentence on a line, the associated eye movements exhibit a very complex trajectory modulated by a large number of variables.

2 Theoretical principles of eye-movement control during reading

Statistics of various types of fixation probabilities and fixation durations serve as benchmark data for all computational models of eye movements during reading. Usually, for each model a number of free parameters is estimated such that summary statistics as those described in Figure 1 are reproduced. The bottom panels of Figure 1 show that our model, called SWIFT, does a good job of recovering fixation probabilities; it does not show that the model also accounts for the differentiated pattern of fixation durations and distributions of landing positions in words contingent on word length and the amplitude of the last saccade (Engbert,

Nuthmann, Richter, and Kliegl, 2005).[2] The model is not unique in this respect – there are at least three other models that can account for such data (McDonald, Carpenter, B., and Shillcock, 2006; Reichle, Pollatsek, Fisher, and Rayner, 1998; Reilly and Radach, 2006; for a comprehensive review of these and other models see Reichle, Pollatsek, and Rayner, 2003). Given this state of the research we can describe how we plan to compare such models in a principled way. But before we turn to the issue of model comparison, evaluation, and development, we describe two of seven core principles that guided the implementation of the SWIFT model: (1) the distinction between when and where to move the eyes and (2) the notion of the perceptual span. These theoretical principles have always been formulated in a qualitative way. Only their implementation in a computer program requires a commitment to a specific mathematical representation.

2.1 When and where to move the eyes

When reading these lines, the control of our eye movements is based on principles about *when* and *where* to launch the next saccade that moves the next word into the fovea for high-acuity analysis. Interestingly, neurophysiological evidence suggests that the temporal and spatial aspects of saccade-generation are largely independent across several levels of organization (Findlay and Walker, 1999).

The temporal aspect of eye-movement behavior, the when decision, is captured by fixation duration measures. Over the past 30 years, much research has been conducted to determine the relationship between fixation durations and linguistic and/or oculomotor variables. First, as described above, it has been shown that various lexical, syntactic, and discourse factors influence fixation durations on words. Thus, fixation durations in reading are sensitive to local processing difficulty (Rayner, 1998). Second, fixation durations are also influenced by low-level nonlinguistic factors. As a consequence, there are fundamental modulations of fixation durations by word length, within-word fixation position, and the distance between fixation locations (i.e., launch site of last saccade), which are unrelated to word recognition (Vitu, McConkie, Kerr, and O'Regan, 2001).

The where pathway, i.e., the question of spatial selection for the next saccade, must solve two tasks: First, which word is to be selected as the target of the next saccade, and, second, what are the principles underlying the control of within-word landing position. The mean value of the landing position distribution is termed the preferred viewing location (Rayner, 1979), which is on average slightly

2 http://www.agnld.uni-potsdam.de/~ralf/swift/.

left of the word center. The initial landing position gives rise to the roughly parabolic refixation probability effect, the optimal viewing position (McConkie, Kerr, Reddix, Zola, and Jacobs, 1989), and the inverted-optimal viewing position (IOVP) for fixation durations (Vitu et al., 2001). The assumption that there are partially independent *when* and *where* pathways (Schad and Engbert, 2012) provides an important boundary condition for the development of psychologically plausible theoretical models of eye-movement control during reading.

2.2 Perceptual span

Analyses of large corpora of eye movements strongly suggest that non-local (distributed) effects of word difficulty on eye fixations during reading are likely to be much more pervasive than suggested by research examining only a few experimentally manipulated target words per sentence (Kliegl et al., 2006; Kliegl, 2007). Distributed processing means that the fixation duration on a word is not only influenced by the characteristics of the fixated word itself but – due to graded parallel word processing within the perceptual span – depends also on the characteristics of the word to the left (lag effect) as well as those of the word to the right of fixation (successor effect). The perceptual span covers an area roughly extending about 15 characters to the right and three characters to the left of the point of fixation in alphabetic languages (McConkie and Rayner, 1975; Rayner, 1975; see Figure 2 for an illustration). Given the decrease of visual acuity with an increase in the eccentricity of information relative to the fixation position, the rate of processing (represented by the height of the functions in Figure 2) is expected to decline. The asymmetry in the direction of reading is interpreted as a modulation of visual perception by attention. Determining which type of information (i.e., visual, sublexical, lexical, semantic) is available from the upcoming word is an area of active and controversial research (see Kliegl et al., 2006; Kliegl, 2007; Rayner, Pollatsek, Drieghe, Slattery, and Reichle, 2007).

Figure 2 also illustrates the proposal that the perceptual span is dynamically modulated by the difficulty of local processing (Schad and Engbert, 2012). Specifically, the peak of processing rate may be lower and the rightward extension may be larger when the eye rests on highly frequent or predictable words in comparison to fixations on low or unpredictable words. Such a dynamical modulation of the perceptual/attentional span implies that processing of a fixated difficult word is processed at a higher rate than on average, but at the cost of reduced parafoveal processing of the upcoming word (Henderson and Ferreira, 1990; Inhoff and Rayner, 1986; Rayner and Pollatsek, 1987).

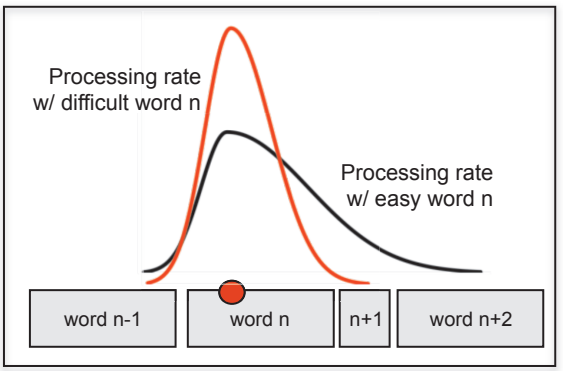

Figure 2: Perceptual/attentional span for a fixation on word n (•) and its dynamical modulation by the difficulty of word *n*. Height of curves represents processing rates for a difficult and an easy word *n*.

2.3 Other core principles

Separate pathways for saccade timing and saccade target selection and spatially distributed processing of an activation field are two of seven core principles of the SWIFT model (Engbert et al., 2005; Table 1). The other five relate to timing of saccade programs and to how this timer can be inhibited with a delay by foveal processing difficulty, to a distinction between different stages in a saccade program, to systematic and random errors in saccade lengths and the implications for mislocated fixations due to such error, and to the relation between saccade latency and saccade amplitude.

3 Computational models of eye-movement control in reading

In Figure 3a we illustrate how the SWIFT model simulates reading the sentence "Manchmal sagen Opfer vor Gericht nicht die volle Wahrheit" [Sometimes victims do not say the complete truth in court]. With time running from top to bottom, the black line indicates the position of the eye at a given moment. Thus, vertical black lines represent fixation durations and horizontal black lines indicate saccades.

The grey hills beneath each word in Figure 3a, show the dynamics of the activation field for the nine words of this sentence in this simulation. Note that some

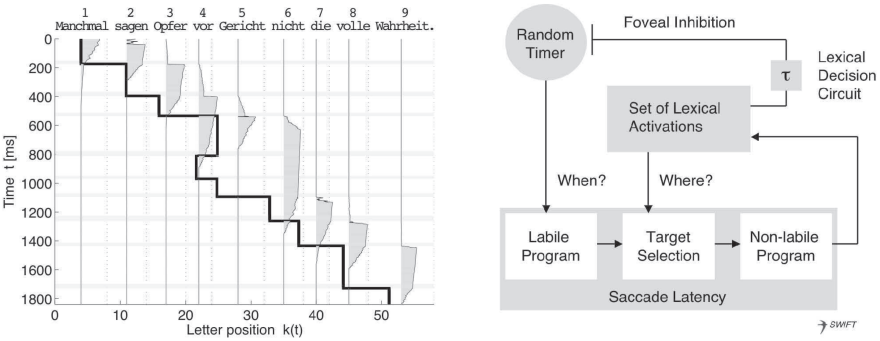

Figure 3: The SWIFT model. Left: Example of a numerical simulation of the SWIFT model Saccade target selection is driven by a spatially-distributed activation field. Word-based activations are illustrated by the shaded areas. The eye's scanpath is indicated by the black line (from Engbert et al., 2005, Fig. 7). Right: Schematic representation of SWIFT. Two independent pathways control fixation duration ("when") and saccade target selection ("where"). A random timer controlling fixation durations can be inhibited to adjust fixation duration to foveal processing difficulty (from Engbert et al., 2005, Fig. 6).

of the activations overlap in time; there is parallel distributed processing. For example, during the first fixation the first three words are active at one point in time. The rate of processing these words depends on how far the word is from the current fixation position. Activation rises steeply for the first, less steeply for the second, and very slowly for the third word. After the first saccade, the rate rises steeply for the third word; also the fourth word is activated because it is now in the perceptual span.

In general, activations related to the N words of a sentence are governed by an N-dimensional set of coupled ordinary differential equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\, a_n(t) = F_n(t)\Lambda_n(t) - \omega \quad (n = 1, 2, 3, \ldots, N), \tag{1}$$

where Λ_n(t) is the processing rate, F_n(t) is a preprocessing factor, which introduces a fast buildup of activation in an early processing stage and is modulated by word predictability p_n, and ω is a global decay process, which we interpret as a memory leakage (see Engbert et al., 2005, for more details of the mathematical formulation of SWIFT).

Turning one such activation "hill" 90° degrees counterclockwise, processing a word means that two random walks spliced at the peak are completed. In SWIFT, word processing difficulty is modulated by printed word frequency and predictability. Low-frequency words have high peaks; high-frequency words have very small peaks. For example, you barely notice the activation associated with

"sich", a reflexive pronoun. Thus, we assume that word frequency, f_n, is related to the maximum of activation, L_n,

$$L_n = \alpha - \beta \log f_n, \tag{2}$$

while predictability, p_n, modulates processing rate. High values of predictability decrease processing rate during parafoveal preprocessing and increase processing rate during lexical completion. Given this dynamics, effects of word length are the consequence of an asymmetric Gaussian-type distribution of processing rate around the current fixation position as shown in Figure 2.

The second principle of separate where-and-when pathways specifies that we must distinguish between target selection (Where?) and timing of saccades (When?). As illustrated in Figure 2b for SWIFT, saccade generation begins with a random timer inducing the start of the next saccade program. The probability to select a word as the next saccade target is computed from its relative lexical activation (i.e., the word's activation value divided by the sum of all lexical activations). Saccade execution occurs only after the necessary saccade-program latency and thereby induces a delay between the effect of lexical activation on target selection and the effect of this saccade on the processing rates in the dynamical field of activations. The set of lexical activations causes also foveal inhibition on the start of the next saccade program (see right panel of Figure 3). Again, this long-loop lexical control process takes time, and this second type of delay is captured within the model parameter τ.

Thus, the movement of the eye depends stochastically on lexical activation of the field of words at the time when this decision is made. Those with high activation are more likely to be selected as saccade targets. For example, for the first saccade of Figure 1, the second and the third word have activation above zero and the second word "won" the competition for being selected as saccade target. Importantly, this single principle of target selection generates all types of eye movements introduced in Figure 1: movement to the next word, skipping, refixations, and regressions. None of the competitor models are as "parsimonious" in this respect.

In the initial versions of SWIFT, processing rate for letters was assumed to follow an asymmetric Gaussian distribution with different parameters, σ_R and σ_L, representing the extension of the span to the right and to the left of the fixation point, respectively. Basically, one of the curves shown in Figure 2 was assumed to apply throughout a simulation. In the dynamical systems framework of SWIFT, discrete processing cycles ("sense" → "think" → "act") are replaced by the temporally continuous evolution of a set of mutually interacting variables representing different cognitive subsystems (Beer, 2000). Within such a framework, it is con-

ceptually possible to implement the dynamic interactions between subsystems very precisely.

We illustrate such a model modification with the proposal of a dynamical span (Schad and Engbert, 2012). When the reading material is difficult, the size of the perceptual span is smaller than for a text of average difficulty (Henderson and Ferreira, 1990). As illustrated in Figure 2, this effect can be accounted for with a sharper distribution of the processing span, determined by parameters σ_R and σ_L, for increasing foveal word difficulty represented by a higher average foveal activation $a_k(t)$ at time t. Specifically, Engbert (2007; see also Schad and Engbert, 2012) assumes that (i) the extension of the processing span to the left is constant; and given by parameter σ_L, (ii) the processing span is symmetric for high foveal load $a_k(t)$; and (iii) the extension to the right, σ_R, increases with decreasing foveal load $a_k(t)$.

This leads to the following relation:

$$\sigma_R = \sigma_L + \delta_1 F(a_k(t)) \tag{3}$$

where δ_1 is a free parameter representing the strength of the dynamical control mechanism and *F(a)* is a sigmoid function.

This is only the beginning. To explore the viability of the concept of a dynamic perceptual span, different mathematical formulations must be implemented and tested by computer simulations and statistical analyses. For example, in ongoing simulations of data from German-English bilingual readers, the Gaussian-type processing span was not constrained enough by data due to its long tail. Therefore, we revised the functional form of the processing span to an inverted quadratic form. As an important property of such a span, we obtained sharp edges of processing. Based on this modification, we were able to estimate the dynamic part of the processing span.

4 Model analysis and comparison

We developed the SWIFT model (Engbert, Longtin, and Kliegl, 2002, Engbert et al., 2005) on the assumption of spatially distributed processing. Another model implementing this assumption is Glenmore (Reilly and Radach, 2003, 2006). In contrast, the E-Z Reader model (Reichle et al., 1998, Reichle, Pollatsek, and Rayner, 2006) is built on the notion of sequential attention shifts (SAS). For the class of SAS models, the serial allocation of visual attention from one word to the next is the central principle driving eye movements. We also contributed to

the SAS-line of research by proposing a model with fewer internal states based on advanced stochastic principles (semi-Markov processes; Engbert and Kliegl, 2001).

Finally, the SERIF model is another fully implemented model which builds upon functional implications of an apparent vertical splitting of the fovea (McDonald et al., 2005).

The development of such quantitative psychological theories is a clear signature of scientific progress. Given the range of competing models of eye-movement control during reading, the need for mathematical analyses and comparisons of models is obvious. Of course, the problem of model selection, analysis, and comparison is a growing area of research in cognitive science in general (e.g., Special Issue on "Model Selection" in the Journal of Mathematical Psychology, 2002).

So how should we compare different mathematical/computational models? Model comparisons are typically based on goodness-of-fit (GOF) statistics, which quantify how much a model's prediction deviates from a given set of experimental data. When using GOF, the underlying assumption is that the model producing the best fit to all data must be a closer approximation to the underlying cognitive process. However, because of random variation in the experimental and statistical methods used (e.g., repeated measurements, inferential statistics), model comparisons based on GOF alone will, in general, produce misleading results (Roberts and Pashler, 2000).

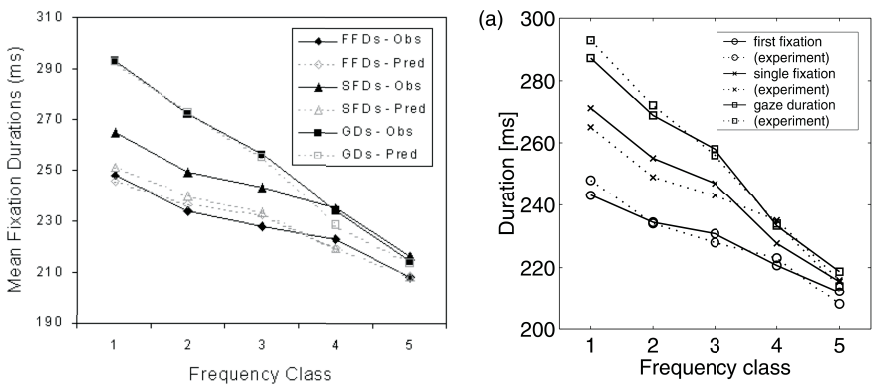

Figure 4: Both (a) the initial SWIFT model (Engbert et al., 2002, Fig. 5a) and (b) the E-Z Reader model (Reichle et al., 2003, Fig. 6 top panel) were fit to the same set of fixation durations measured for 5 different word-frequency classes (data from Schilling et al., 1998). Lines refer to observed (Obs) and predicted (Pred) gaze durations (GDs), single fixation durations (SFDs), and first fixation durations (FFDs).

Moreover, advanced mathematical models often fit experimental data equally well. For example, SWIFT (Engbert et al., 2002) and E-Z Reader (Reichle et al., 2006) reproduced fixation durations reported in Schilling, Rayner, and Chumbley (1998) equally well in terms of GOF (see Figure 4). The lines represent observed and predicted gaze durations (i.e., the sum of fixations when a word is first read), single-fixation durations (i.e., when a word is fixated exactly once), and first-fixation durations (i.e., the duration of the first fixation, irrespective of how many fixations occurred) as a function of five log-frequency classes.

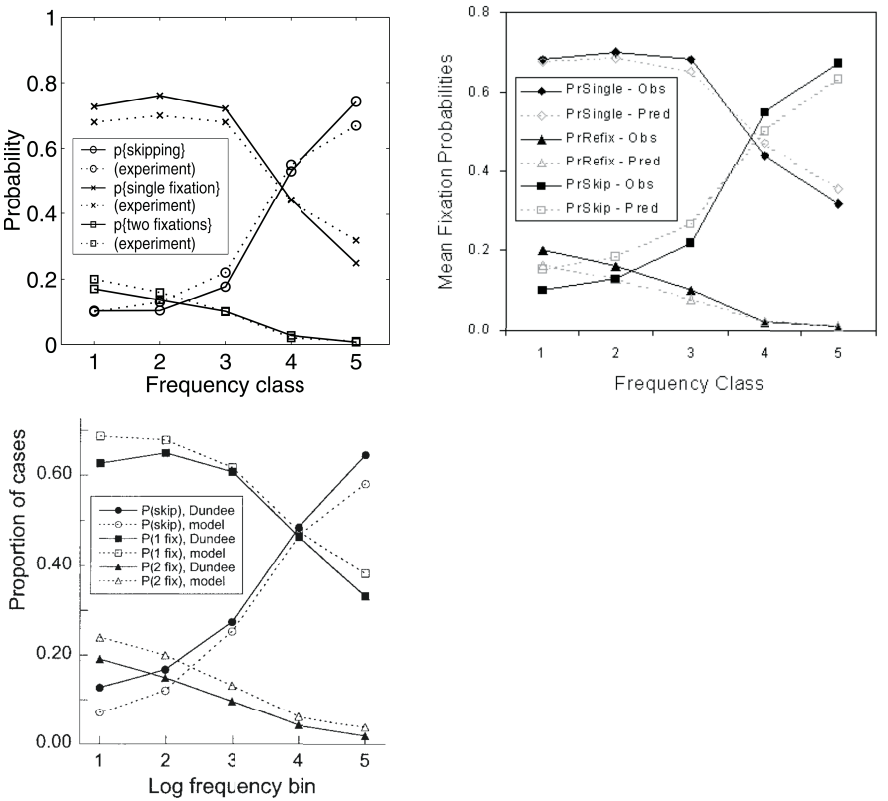

Figure 5: Data and simulations of single-fixation, skipping, and refixation probabilities as a function of log10 word frequency (per million words) for the E-Z Reader model (left, Reichle et al., 2003, Fig. 6 bottom panel) and the initial SWIFT model (middle, Engbert et al., 2002, Fig. 5b) were fit to the same set of fixation durations measured for 5 different word-frequency classes (data from English sentences; Schilling et al., 1998). The right panel shows the fit of the SERIF model (McDonald et al., 2005, Fig. 3b; data from English newspaper texts [Dundee corpus], Kennedy, 2003).

Figure 5 (left and middle panel) illustrates the similarity in fit relating to probabilities for single fixations, skipping, or refixations as a function of word frequency for the same data (Schilling et al., 1998) and the same two models (E-Reader 9, Reichle et al., 2006; SWIFT, Engbert et al., 2002). Interestingly, the simulations of the SERIF model (McDonald et al., 2005) led to a very similar pattern of means although the model was fit to a completely different set of English eye movement data (i.e., Dundee Corpus; Kennedy 2003).[3]

We consider the qualitative (and also largely quantitative) agreement between observations and simulations as quite remarkable. Therefore, to restate the argument from above, for this research field at this point in time, the qualitative agreement between model and data is more important than differences in quantitative goodness-of-fit statistics (such as root mean squared error).

If we grant comparability in goodness of fit, we may still compare the models with respect to their complexity. One indicator of model complexity is the number of free model parameters. Interestingly, the models do not differ on this dimension either, ranging from 13 to 18. Moreover, many of the parameters are only free to vary within a range dictated by substantive issues. For example, in the SWIFT model, the asymmetry of the span, estimated as Gaussian standard deviations for the left and right processing-rate functions (see Figure 2), must map onto a plausible range of letters (i.e., 3 to the left and around 10 to the right). Similarly, parameters estimating delay lines in the model are narrowly constrained by the physiology of the structures known to be involved in saccade programming and execution.

The dissimilarity of models, immediately apparent from the cartoons in Figure 6, can be traced to fundamental decisions about their architectures (E-Z Reader: stochastic automaton, SWIFT: stochastic dynamical system, SERIF: stochastic model, Glenmore: connectionist network/coupled difference equations) and to their implementation of temporal (when) and spatial (where) decisions. The fact that the models do a comparably good job in accounting for benchmark results with a comparable degree of model complexity as indexed by the number and constraints on free parameters, strongly suggests that the models are not sufficiently constrained by the benchmark results. Therefore, a more promising route than comparisons in terms of goodness of fit, is to increase the scope of results they are expected to simulate.

3 Differences between Figure 5 and Figure 1 (which displays similar probabilities) are due to language differences between English and German. Data in Figure 1 were fit by the later version of SWIFT (Engbert et al., 2005).

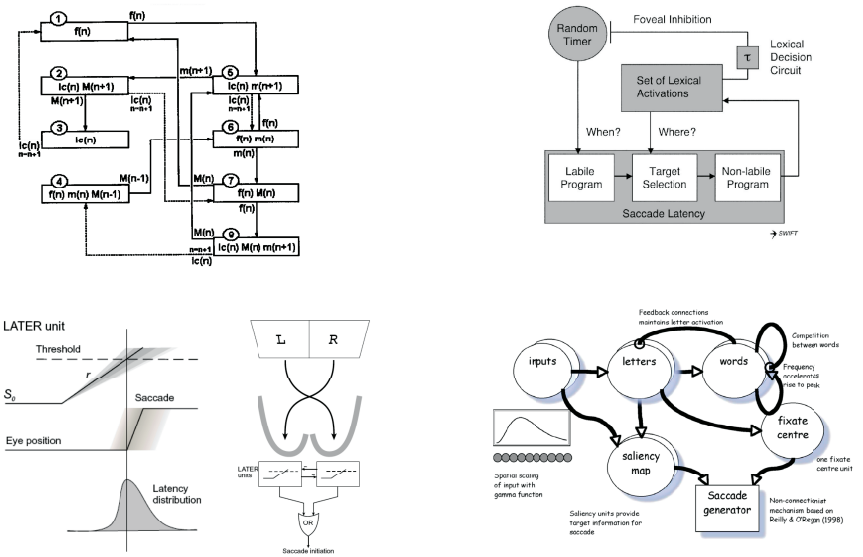

Figure 6: Architectural blueprints of E-Z Reader (top left, from Reichle et al., 1998, Fig. 4), SWIFT (top right, from Engbert et al., 2005, Fig. 6), SERIF (bottom left, from McDonald et al., 2005, Fig. 1), and Glenmore (bottom right, from Reilly and Radach, 2006, Fig. 1) exhibit a high dissimilarity between these computational models of eye-movement control during reading.

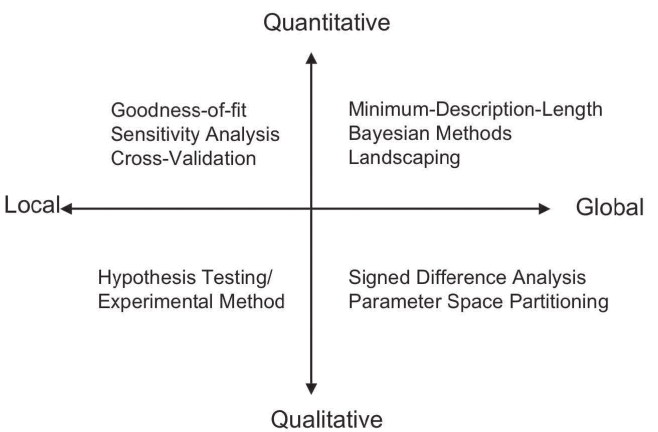

Figure 7: Classification of methods for model analysis and comparison in a two-dimensional space (modified from Pitt et al., 2006), defined by the degree to which the method evaluates quantitative versus qualitative model performance (vertical axis) and whether the method focuses on local or global model behavior (horizontal axis).

Pitt, Kim, Navarro, and Myung (2006) suggest a classification of methods for model analysis and selection in a two-dimensional space, where the first dimension represents the range from local to global methods and the second dimension is a scale from qualitative methods to quantitative methods (Figure 7).

Psychological research almost exclusively applies local methods, but methods representing the global approach are needed to test model reliability and generalizability. The reason for this imbalance is that the applicability of quantitative global methods to the diverse range of models in psychology is currently "limited by their technical requirements" (Pitt et al., 2006). An important reason is that realistic models are computational rather than analytical, which creates problems for the implementation of methods of model analysis and comparison.

5 Meeting the Roberts and Pashler (2000) challenge

Global methods for data analysis, quantitative and qualitative, may become available in a convincing way for psychological models over the next years. Currently, we do not see a straightforward application for the models in our domain of research. In psychology, the problem is that we are usually satisfied with good model fits, rarely complemented by sensitivity analyses of parameters or cross-validation of results. Therefore, for now, we may remain within the local-quantitative quadrant of Figure 7. Roberts and Pashler's (2000, 2002) starting point is that, although psychological models may live up to reasonable expectations about goodness of fit, goodness of fit does not discriminate convincingly between the models and is only a necessary condition for model evaluation. Consequently, they argued that there is currently no psychological theory in the sense in which we use this term in physics because psychological theories (or models) fall short on the following three criteria: strictness of model, reliability of data, and unexpected predictions. True or not, let us proceed from the assumption that several computational models of eye-movement control in reading, varying widely in theoretical assumptions and architectures, recover critical experimental benchmark results with the same number of free parameters, but cannot be distinguished in goodness of fit. In the following we review research and outline a research program that, at least in perspective, may allow us to meet this challenge for the SWIFT model.

5.1 Strictness of model and reliability of data

Figure 8 serves to illustrate the first two criteria: strictness of model and reliability of data. The left panel is taken from Roberts and Pashler (2000). Crosses with error bars represent two experimental measures or estimates A and B that can be derived from empirical data. Depending on the reliability of the experimental data, the error bars will be narrow (left column) or wide (right column). The grey points are results from model simulations; they are predictions of the models for different combinations of the model parameters. Thus, a strict model (top row) will generate a smaller set of predictions than a very flexible model (bottom row). In each panel, the data fall into the grey model zone. Thus, the models are always consistent with the data. By itself, however, the goodness-of-fit criterion does not say anything about model strictness and data reliability. The models in the bottom row are too flexible and the data in the right column are too variable. Strong support of the model is only present in the top left panel, where model strictness and reliability of data are in a reasonable relation with each other.

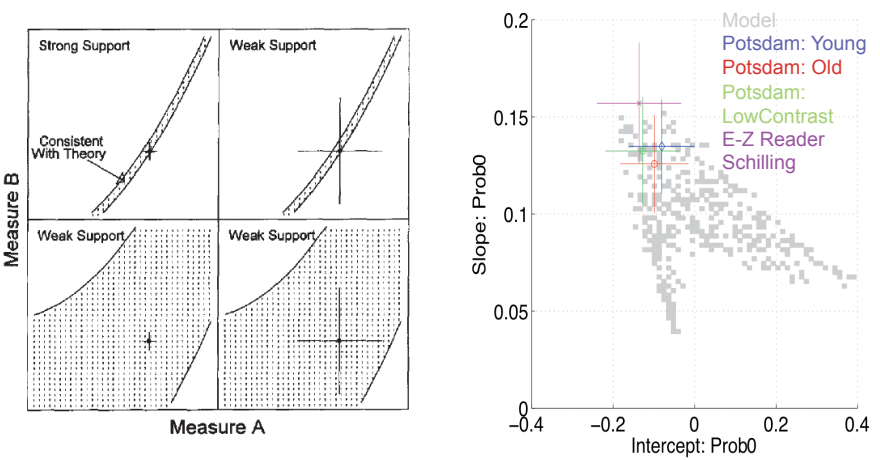

Figure 8: Left: Illustration of model strictness and reliability of data (from Roberts and Pashler, 2000, Fig. 1). Crosses represent experimental results; grey dots represent model simulations. Criteria for strict model and reliable data fulfilled for top-left panel. Right: Application to relation between intercept (x-axis) and slope (y-axs) for regression of skipping probability on log of word frequency. Crosses indicate different data sets; grey dots indicate results from SWIFT simulations.

As a concrete example from our research (see right panel of Figure 8), we regressed skipping probability on log of word frequency; from this we obtain two "mea-

sures", the intercept and the slope (which is positive as skipping increases with word frequency). In a next step we carry out a large number of model simulations with parameters drawn randomly from reasonable ranges. From the data of each simulation we derive intercept and slope for the two measures. These values, again, are shown as grey dots and represent possible predictions by the SWIFT model. The crosses represent intercepts and slopes from different data sets, comprising young and old German readers, young German readers reading with strongly reduced screen contrast, and young English readers. These results show that the reliability of the estimates is quite comparable across data sets and more importantly that the between-language variation has a much stronger effect than the within-language age or within-language contrast manipulation. This suggests that language-comparative research in reading may hold much potential to explore the "legal" parameter settings of the model. Most importantly, however, we argue that these results suggest that the between-simulation variability of the SWIFT model is in a reasonable agreement with the experimental results. The next step is to expand the measurement space and, of course, to engage in systematic comparisons between models of claiming similar goodness of fit of benchmark results with similar number of model parameters.

5.2 Unexpected predictions

By far the toughest criterion to meet, the gold standard for a model is to generate predictions about behavior that is subsequently recovered from the data or experimentally established. Lakatos (1978, 6) put this succinctly:

> The hallmark of empirical progress is not trivial verification. ... It is no success for Newtonian theory that stones, when dropped, fall towards the earth, no matter how often this is repeated. But so-called 'refutations' are not the hallmark of empirical failure, as Popper has preached, since all programmes grow in a permanent ocean of anomalies. What really counts are dramatic, unexpected stunning predictions: a few of them are enough to tilt the balance...

As an example for a stunning prediction of Newtonian physics, he mentions Halley's exact prediction of space and time for the return of Halley's comet 72 years later. Lakatos expresses the essence of the third problem: The a priori probability that the theory will fit the data is often ignored. At the end of this section we will give an example from our research, which was surprising to us. This example, however, is not representative of normal model development, given that none of the psychological theories we are aware of would seriously claim to be in the Newtonian league of scientific theories. Indeed, history of science regularly

uncovers the meandering between alternative conceptualizations and difficulties in choosing between them at the time of emergence of theories, which are now known only in a single canonical form.[4]

5.2.1 Case 1: Surprising results, incompatible with model predictions

Most frequently, development of psychological models is driven by new experimental results. Of course, in part this is simply due to the fact that there are many more experimental psychologists contributing new knowledge than there are modelers (and models) who can devote time to address the new results with their models. Indeed, we suspect that models will often not handle new experimental results adequately as they are implemented at the time of their publication. However, the results may not be incompatible with the theoretical principles guiding the model implementation (Rayner, 2009).

One example of such a result relates to fixation durations prior to skipped words. Experimentally, we observed that fixations before skipped words were shorter before short (or high-frequency) words ("skipping benefit") and longer before long (or low-frequency) words ("skipping cost"; see Kliegl and Engbert, 2005, Kliegl, 2007, for details). The observation of skipping benefits is critical for models based on sequential attention shifts (SAS) like E-Z Reader (Reichle et al., 1998). In such a model, word skipping can only be produced by (i) cancellation of a saccade program to the next word n+1 and (ii) the initiation of a new saccade program to word n+2. As a consequence, models of the SAS class always generate skipping costs, i.e., longer fixation duration before skipped words.

The SWIFT model (Engbert et al., 2005) also predicts skipping cost but this prediction was not tied as strongly to the model architecture as it is for the E-Z Reader model. The example is quite illustrative, because the skipping cost arises for a very different reason: the longer a fixation duration, the longer the preprocessing of the next word, and the higher the chances that the next word will be skipped. Thus, whereas in E-Z Reader long fixations prior to skipped words are a consequence of skipping, they are the cause of skipping in SWIFT. Nevertheless, neither model correctly recovered the skipping benefit associated with short words – counter to our own experimental data. Fortunately, in psychology, such reports of a falsification of a model do not necessarily preclude publication.

---

4 For example, Damerow, Freudenthal, McLauhglin, and Renn (1992) describe and explain the problems and misunderstandings during the transition from early concepts of motion to the theory of motion in classical mechanics, using, among others, texts by Descartes and Galileo about the free fall of bodies and the composition of motions and forces.

5.2.2 Case 2: Surprising results, compatible with model after its modification

Box's (1979) "all models are wrong, some models are useful" is the guiding overarching principle. Falsifications are useful if they inspire model modification that encompasses new results in a principled way, rather than by some ad-hoc fix of the model. Indeed, such results frequently spur model modification to account for the results in such a way that ideally previous successful simulation results are preserved. Since the original publication of the failure to account for skipping cost, we have used this failure as one starting point for the further development of the model. In particular, we implemented the theoretical proposal of a dynamical modulation of the perceptual span, contingent on the foveal processing difficulty as described in section 2.2 (see also Figure 2, Eq. 3). We assumed that (i) the extension of the processing span to the left is constant, (ii) the processing span is symmetric for high foveal load, and (iii) the extension to the right increases with decreasing foveal load. Next, we fitted all model parameters of this variant of the SWIFT model using the same methods as reported by Engbert et al. (2005).

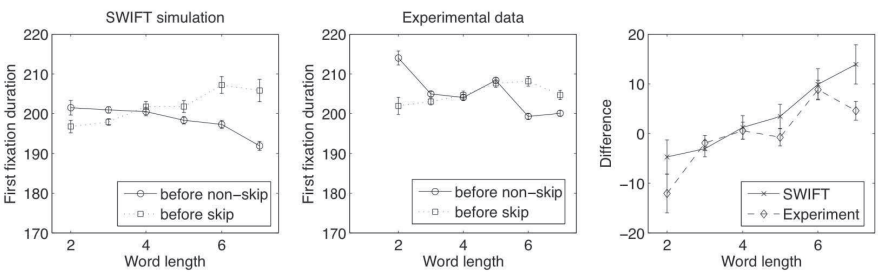

Figure 9: Skipping costs and benefits as a function of word length in experimental data and SWIFT simulations. The left panel shows average fixations durations in SWIFT simulations. The center panel shows the same plot for experimental data. The right panel shows the fixation durations before skipping subtracted by the fixation durations before non-skipping as a function of word length. The model simulations reproduce the experimental result that there are skipping costs for long words (word length > 5 letters) and skipping benefits for short words (< 4 letters).

Are there consequences of the dynamic processing span in SWIFT for the issue of word skipping discussed in the last section? Specifically, will this modification reveal skipping benefit prior to short words? Since parafoveal processing is very important to word skipping, it is plausible that the dynamical modulation of the perceptual span will lead to new results. Experimentally, we had observed that fixations before skipped words were shorter before short (or high-frequency)

words and longer before long (or low-frequency) words (Figure 9, center panel; see Kliegl and Engbert, 2005, for details). Interestingly, the SWIFT variant with dynamic foveal span can reproduce this highly specific data pattern accurately (Figure 9, left panel). The good agreement between experimental data and SWIFT simulations can be made visible, when differences in fixation durations (fixation durations before skippings subtracted by fixation durations before non-skipping) are plotted (Figure 9, right panel).

These results from pilot simulations represent a major model improvement, because the current version of SWIFT always generated skipping costs (increased fixation durations before skipping) between 10 ms (word length 2) and 60 ms (word length 6). To our knowledge, the variant of the SWIFT model with dynamic foveal processing span investigated here is the only computational model that can reproduce the patterns of fixation durations before skipped words.

The model modification also "survived" two important tests. First, adding a new principle to an existing model might change the model's performance on benchmark tests. Evaluations based on summary statistics for fixation durations and fixation probabilities, however, indicate that the dynamic processing span is as compatible with experimental data as a constant, asymmetric processing span. The overall goodness-of-fit of the model was not affected by the dynamic processing span. Second, the introduction of a dynamic processing span might have a strong impact on the effects of word properties of the last and next words (Kliegl et al., 2006), because variations of the extension of the processing span in general will change parafoveal processing. Interestingly, such changes in model performance were not observed from our pilot simulations. Thus, at this point the dynamical span served as defensible extension of the original SWIFT model. For a continuation of this story in the context of a further modification of the model we refer to Engbert and Kliegl (2011). We submit that this back and forth between experimental results and model development accounts for most of the research time in model development.

5.2.3 Case 3 (Lakatos Case): An unexpected prediction lurking in the data

We conclude with a "gold standard" example of an unexpected prediction derived from the SWIFT model and confirmed by the analyses of data collected many years ahead of the prediction (Risse et al., 2008): *Refixation probability should be larger after a skipped function word, but not after a skipped content word.*

The prediction is derived from the assumption that there is a competition between saccade targets. Suppose you are looking at a word and next to this word is a preposition. The preposition will likely be processed during this fixation,

because prepositions are among the most frequent and most predictable words. This implies that the next word drops out of the race to become selected as the next saccade target. At the same time the second word to the right will be slowly raising its activity level. So its chances to become elected increase compared to the situation with a competitor in position N+1. And most importantly, it will be fixated early in its activation profile, which increases the chances that it will be refixated.

The results are presented in Figure 10 and match the prediction very well. After a skipped function word the refixation probability is higher than after a skipped content word. There is some misfit, too: The overall refixation rate after skipping is overestimated.

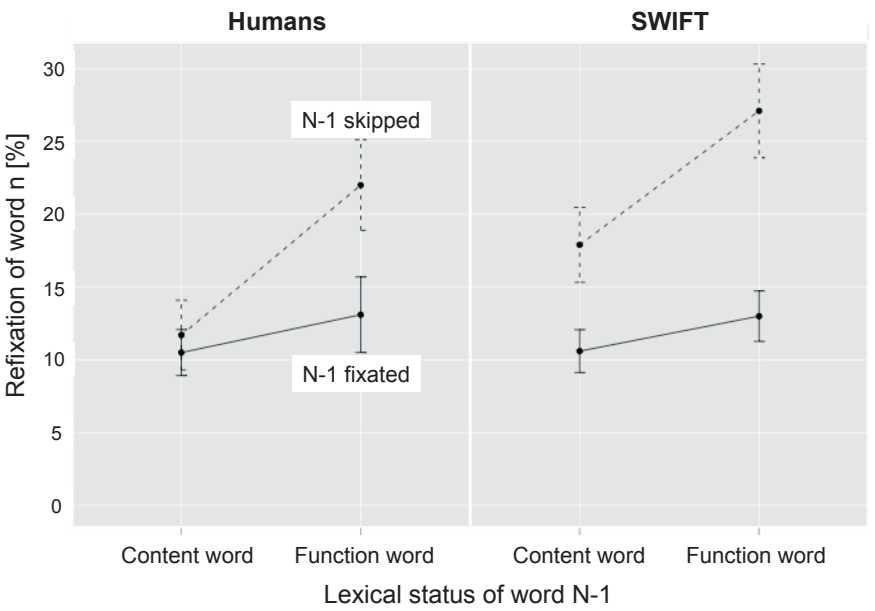

Figure 10: Refixation rate after skipped and fixated content and function words. Left: Human data. Right: SWIFT simulation results.

Finally, the prediction lends itself to a comparison with the E-Z Reader model which also makes the general prediction that refixation rate will be higher after skipped words, but will generate the opposite prediction with respect to lexical status: Refixation rate should be *lower* after a skipped function word than a skipped content word. The reason for this prediction is that a skipped function

word provides longer preview of the following word than a skipped content word in this model. So there will be more need for processing after a skipped content word than after a skipped function word.

6 Perspectives and conclusion

6.1 Implications beyond reading

Experimental and mathematical psychology have developed detailed models of the interplay between cognitive subsystems (e.g., perception, attention, language, motor control). Dynamic models based on this approach can provide powerful theoretical blueprints for the behavioral and also for the neural organization of these cognitive processes,

(1) if they are simulated on a computer with advanced techniques and studied qualitatively within the framework of nonlinear-science models,
(2) if model parameters are estimated from high-resolution time series, and
(3) if both experimental data and model simulations are evaluated by advanced methods for the analysis of complex multivariate time series.

As an example, we have described competing models that aspire to meet these criteria in the domain of reading research (Engbert et al., 2005; McDonald et al., 2005; Reichle et al., 1998; Reilly and Radach, 2006). There are hardly any more convenient measures than eye movements if one is interested in how behavior rapidly unfolds over time. Thus, eye movements represent an ideal model system in experimental psychology.

Most importantly, the neural circuits subserving the generation of eye movements are well understood (Sparks, 2002). Eye movements may well be the most direct behavioral signatures of neural firing, for they are directly related to spatio-temporal activation in the superior colliculus (SC). Indeed, the minimum of the oculomotor response time is about 60 ms after visual stimulus presentation, where the estimate is based on brainstem circuitry. More and more is currently learned about how higher-order structures (e.g., frontal eye fields, lateral intraparietal cortex, visual cortex) modulate brainstem nuclei when the oculomotor system is triggered by perceptual, attentional, and vestibular demands (e.g., Munoz and Everling, 2004) Further, because the loads on the extraocular muscles do not vary, reverse modeling can be used to reconstruct the eliciting innervation pattern. Most importantly, saccade and fixation parameters describing the eye movements across a visual scene or across a text embody behavioral dynamics

in experimental designs covering the broad spectrum of behavioral activity from simple perception via reading to postural control.

6.2 Model analysis and comparison

Starting with Huey (1908), research of eye movements in reading has been impressed with the range and stability of differences between individual readers as well as the magnitude of effects induced by differences in task demands. For example, individual differences in single-fixation durations among readers varying from 18 to 80 years of age account for more variance than 18 fixation-positions and psycholinguistic predictors (Kliegl et al., 2006). In agreement with the early research, preliminary analyses of data from bilingual readers of English and German varying widely in second-language proficiency suggest that individual differences will be even more pronounced in skipping and refixation probabilities. Thus, accounting for this variance in the SWIFT model would represent a major step in the further development of the model. We also note that there is no other computational model of reading or other cognitive processes that has been expanded in this direction.

Summary statistics relating to fixation durations and probabilities as a function of word length and word frequency can be reproduced remarkably well by at least four computational models of eye-movement control during reading. They all succeed with respect to the necessary condition of **goodness of fit** with a comparable number of free model parameters. Here we went beyond this necessary condition and offer some evidence that the SWIFT model may also live up to expectations of a strict set of criteria relating to model strictness, reliability of data, and unexpected predictions, as postulated by Roberts and Pashler (2000).

As a test of **model strictness** and **reliability of the data** we showed that the covariation of intercept and slope from the regression of word-skipping probability on log word frequency across simulations of the SWIFT model with random variation of model parameters within plausible ranges of parameter values agrees very well with variation observed between different reader groups varying in age, contrast of screen, and language.

The requirement of **unexpected model predictions** is illustrated in the form of three cases. First, in psychological research it is still more common to be surprised by new results. They may be compatible with model principles, but not recovered by a model in its current implementation. Second, some surprising results are bound to lead to constructive modifications of principles and implementations. Usually, these new “successes” are to be cumulative to earlier ones. Third, sometimes model predictions can be evaluated with respect to their

agreement with previously not known facts. Such predictions should have a low a priori probability in the scientific community; they must not be trivial.

Model modification usually changes model complexity and requires a consideration of its own. Neal (1996, 103–104) aptly summarized the issue of model complexity as follows:

> Sometimes a simple model will outperform a more complex model ... Nevertheless, ... deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well.

In summary, we have contributed key findings to both experimental and computational aspects of eye-movement control during reading. We developed a computational model based on the assumption on distributed processing (SWIFT; Engbert, et al., 2002, 2005, Schad and Engbert, 2012). The model accounts for a large number of experimental observations, e.g., various measures of inspection probabilities and inspection durations, eye landing positions within words, delayed lexical access, parafoveal preprocessing. With respect to the scope of covered phenomena and transparency of its theoretical principles, arguably, SWIFT is currently the most advanced model of eye-movement control during reading. Of course, there are also aspects of reading behavior the model cannot and cannot be expected to get right at this point in time (see Engbert et al., 2005; Risse et al., 2008), but it certainly is a very useful tool guiding much of our research (Box, 1979).

6.3 Conclusion

We like to think about eye movements during reading as the "drosophila of psychological modeling" because they map onto a comparatively simple measurement space within which behavior of a surprisingly high level of complexity unfolds. It is a general critical requirement for modeling of cognitive processes to focus on a field of study with just the right level of complexity of behavior for the intended model. Eye movements during reading appear to meet this expectation in an ideal way.

7 Acknowledgement

This research was supported by Deutsche Forschungsgemeinschaft Research Group 868 "Computational Modeling of Behavioral, Cognitive, and Neural Dynamics" (Grant KL955/14) and European Science Foundation (ESF; Grant 05_ECRP_FP_006, DFG KL955/7).

References

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* 4. 91–99.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In: Launer, R. L. & Wilkinson, G. N. (eds.), *Robustness in statistics*. New York: Academic Press.

Damerow, P., Freudenthal, G., McLaughlin, P., & Renn, J. (1992). *Exploring the limits of preclassical mechanics*. New York: Springer.

Engbert, R. (2007). *Reading with a dynamic processing span*. Presentation at 14th ECEM, Potsdam, Germany.

Engbert, R. & Kliegl, R. (2011). Parallel graded attention models in reading. In: Liversedge, S. P., Gilchrist, I. & Everling, S. (eds.). *The Oxford Handbook of Eye-Movement*s. New York, NY: Oxford University Press. 787–800.

Engbert, R. & Kliegl, R. (2001). Mathematical models of eye movements in reading: A possible role for autonomous saccades. *Biological Cybernetics* 85. 77–87.

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

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

Findlay, J. M. & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences* 22. 661–721.

Henderson, J. M. & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16. 417–429.

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

Inhoff, A. W. & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics* 40. 431–439.

Kennedy, A. (2003). The Dundee corpus [CD-ROM]. Dundee, Scotland: University of Dundee, Department of Psychology.

Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, & Reichle. *Journal of Experimental Psychology: General* 136. 530–537.

Kliegl, R. & Engbert, R. (2005). Fixation durations before word skipping in reading. *Psychonomic Bulletin & Review* 12. 132–138.

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. 262–284.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General* 135. 12–35.

Lakatos. I. (1978). The methodology of scientific research programmes. In: Worrall, J. & Currie, G. (eds.). *Philosophical Papers* 1. Cambridge: Cambridge University Press.

McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., & Jacobs, A. M. (1989). Eye movement control during reading: II. Frequency of refixating a word. *Perception & Psychophysics* 46. 245–253.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics* 17. 578–586.

McDonald, S. A., Carpenter, R. H. S., & Shillcock, R. C. (2005). An anatomically constrained, stochastic model of eye movement control in reading. *Psychological Review* 112. 814–840.

Munoz, D. P. & Everling, S. (2004). Look away: the anti-saccade task and the voluntary control of eye movement. *Nature Review Neuroscience* 5. 218–228.

Neal, R. M. (1996) *Bayesian learning for neural networks*. New York: Springer.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review* 113. 57–83.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology* 62. 1457–1506.

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

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception* 8. 21–30.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology* 7(1). 65–81.

Rayner, K. & Pollatsek, A. (1987). Eye movements in reading: A tutorial review. In: Coltheart, M. (ed.). *Attention and Performance* 12. New York: Academic Press.

Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T. J., & Reichle, E. D. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert (2006). *Journal of Experimental Psychology: General* 136. 520–529.

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

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

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences* 26. 446–526.

Reilly, R. & Radach, R. (2003). Foundations of an interactive activation model of eye movement control in reading. In: Hyönä, J., Radach, R., & Deubel, H. (eds.). *The mind's eye: Cognition and applied aspects of eye movement research*. Oxford: Elsevier. 429–455.

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

Risse, S., Engbert, R., & Kliegl, R. (2008). Eye-movement control in reading: Experimental and corpus-analytic challenges for a computational model. In: Rayner, K., Shen, D., Bai, X.,

& Yan, G. (eds.). *Cognitive and cultural influences on eye movements*. Tianjin: Tianjin People's Publishing House. 65–91.
Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107. 358–367.
Roberts, S. & Pashler, H. (2002). Reply to Roders and Rowe. *Psychological Review* 109. 605–607.
Schad, D. J. & Engbert, R. (2012). The zoom lens of attention: Simulating shuffled versus normal text reading using the SWIFT model. *Visual Cognition*, 20(4–5). 391–421.
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. 1270–1281.
Sparks, D. L. (2002). The brainstem control of saccadic eye movements. *Nature Review Neuroscience* 3. 952–964.
Vitu, F., McConkie, G. W., Kerr, P., & O'Regan, J. K. (2001). Fixation location effects on fixation durations during reading: an inverted optimal viewing position effect. *Vision Research* 41(2526). 3513–3533.

Prof. Dr. Reinhold Kliegl
University of Potsdam
Department of Psychology
Karl-Liebknecht-Str. 24–25
14476 Potsdam
Germany
kliegl@uni-potsdam.de

Prof. Dr. Ralf Engbert
University of Potsdam
Department of Psychology
Karl-Liebknecht-Str. 24–25
14476 Potsdam
Germany
ralf.engbert@uni-potsdam.de